The value of publicly available, textual and non-textual information for startup performance prediction

Kaiser, Ulrich ; Kuhn, Johan M

Abstract: We use administrative textual and non-textual data retrieved from publicly available archives to predict the performance of Danish startups at the time of foundation. The performance outcomes we consider are survival, high employment growth, a return on assets of above 20 percent, new patent applications and participation in an innovation subsidy program. We consider a base specification that includes variables for legal form, region, ownership and industry in all specifications and add variable sets representing firm names, business purpose statements (BPSs) as well as founder and startup characteristics. To forecast the two innovation-related performance outcomes well, we only need to include a set of variables derived from the BPS texts on top of the base variables while an accurate prediction of startup survival requires the combination of the firm names and the BPS variables along with founder characteristics. An accurate forecast of high employment growth needs the combination of the BPS variables and the founder characteristics. All information our forecasts require is likely to be easily obtainable since the underlying information is mandatory to report upon business registration in many countries. The substantial accuracy of our predictions for survival, employment growth, new patents and participation in innovation subsidy programs indicates ample scope for algorithmic scoring models as an additional pillar of funding and innovation support decisions.

DOI: https://doi.org/10.1016/j.jbvi.2020.e00179

Journal of Business Venturing Insights 14 (2020) e00179

Contents lists available at ScienceDirect

Journal of Business Venturing Insights

journal homepage: www.elsevier.com/locate/jbvi

The value of publicly available, textual and non-textual information for startup performance prediction

Ulrich Kaiser [a,b,c,d,*], Johan M. Kuhn [b,e]

[a] University of Zurich, Department of Business Administration Chair for Entrepreneurship, Plattenstrasse 14 CH-8032, Zurich, Switzerland
[b] Copenhagen Business School, Department of Strategy and Innovation, Denmark
[c] ZEW Leibniz Center for European Economic Research, Mannheim, Germany
[d] Institute for the Study of Labor, Bonn, Germany
[e] EPAC, Economic Policy, Analysis & Consulting, Jellebakken 1, 8240 Risskov, Denmark

ARTICLE INFO

JEL classification:
L26
C53

Keywords:
Startup
Performance
Prediction
Text as data
Algorithmic scoring

ABSTRACT

We use administrative textual and non-textual data retrieved from publicly available archives to predict the performance of Danish startups at the time of foundation. The performance outcomes we consider are survival, high employment growth, a return on assets of above 20 percent, new patent applications and participation in an innovation subsidy program. We consider a base specification that includes variables for legal form, region, ownership and industry in all specifications and add variable sets representing firm names, business purpose statements (BPSs) as well as founder and startup characteristics. To forecast the two innovation-related performance outcomes well, we only need to include a set of variables derived from the BPS texts on top of the base variables while an accurate prediction of startup survival requires the combination of the firm names and the BPS variables along with founder characteristics. An accurate forecast of high employment growth needs the combination of the BPS variables and the founder characteristics. All information our forecasts require is likely to be easily obtainable since the underlying information is mandatory to report upon business registration in many countries. The substantial accuracy of our predictions for survival, employment growth, new patents and participation in innovation subsidy programs indicates ample scope for algorithmic scoring models as an additional pillar of funding and innovation support decisions.

1. Introduction

Identifying promising startups is a formidable task for investors, creditors and policy makers alike. Even though each group of stakeholders often has quite a wealth of information available when deciding about a possible involvement in a particular startup, this information must be processed quickly which in turn implies that simple heuristics become highly valuable (Baum and Wally, 2003; Eisenhardt, 1989; Kirsch et al., 2009). In addition, investors and creditors aim at identifying promisinging startups early and therefore increasingly often use algorithmic scoring models (Corea, 2018; Diffey, 2019; Palmer, 2017). More generally, uncertainties in the ex-ante evaluation of business opportunities are fundamental to the theory and the empirical testing of entrepreneurial strategy (Ahuja et al., 2005; Amit et al., 1990; Dencker and Gruber, 2015; Nikiforou et al., 2019; Oriani and Sobrero, 2008).

* Corresponding author. University of Zurich, Department of Business Administration, Chair for Entrepreneurship, Plattenstrasse 14, CH-8032, Zurich, Switzerland.
E-mail addresses: ulrich.kaiser@business.uzh.ch (U. Kaiser), johan@epacn.dk (J.M. Kuhn).

https://doi.org/10.1016/j.jbvi.2020.e00179
Received 13 March 2020; Received in revised form 2 June 2020; Accepted 5 June 2020

We put ourselves in the shoes of key stakeholders in startups — investors, creditors and policy makers — and ask if it is possible to accurately predict the expected performance of a focal startup using high quality, publicly available administrative data and simple econometric methods. Given that the corporate and the academic world has seen a surge in the availability of such data (Card et al., 2010; Einav and Levin, 2013; Gentzkow et al., 2019), we see great potential for better informing stakeholders by making use of public data treasures, perhaps even to a degree that allows for algorithmic scoring models. A leading example for the increased availability of public data are the US where websites like https://www.data.gov/maintained by the US government and https://fred.stlouisfed.org/ maintained by the Federal Reserve Bank of St. Louis are already making administrative data available online. It seems likely that open access to such data will grow further with as many as 79 countries worldwide having joined the "Open Government Partnership" initiative with its explicit goal to ease access to public data (https://www.opengovpartnership.org/). The Danish administrative data we use in this study in fact have their origins in this initiative that the country joined in 2011.[1]

We consider five different performance proxies that are relevant to different types of stakeholders. Our first performance measure is survival which is a variable that all stakeholders have an inherent interest in, most notably banks which mostly worry about default probabilities. Our second and third performance measures, high employment growth and high growth in the returns on assets, are of high importance for investors while our fourth and fifth performance measure, successful participation in a government innovation subsidy program after foundation and new patents, are closely watched by policy makers.

The set of performance predictors we consider comprises of a set of "baseline" variables that have found wide applicability in entrepreneurship research, namely legal form, region, ownership, founder gender, and industry, as well as sets of (i) firm name variables, (ii) variables generated from firms' business purpose statements (BPSs or "articles of organization", "articles of incorporation" or "certificate of incorporation") which are mandatory for corporations worldwide and required by most US states and most European countries as an integral part of their business formation documents, (iii) basic founder characteristics like previous founding experience as well as (iv) initial startup characteristics like initial assets and profits along with address information.

We base our analysis on the population of Danish firms started as incorporated companies between 2012 and 2014, 55914 firms in total, whose data we downloaded via the open APIs of Danish government websites. We run simple simple logit models for each of our five performance models and assess the forecasting accuracy of our specifications by calculating the respective areas under the receiver-operator curves (AUCs). The AUC is a frequently applied, scale-free forecast performance statistic for binary firm performance models (Agrawal and Taffler, 2008; Åstebro and Winter, 2012; Chava and Jarrow, 2004). We analyze the contribution of each individual set of explanatory variables to forecasting precision since not all data may be publicly available in all countries and since there are differences in their ease of use.

Our key findings are that (i) our models predict all performance outcomes with high accuracy, with the exception of high return on assets, (ii) the data needed to generate precise forecasts are both easily obtainable and straightforward to apply in simple empirical models and (iii) "text-as-data" (Catalini et al. 2019; Gentzkow et al., 2019; Guzman and Stern, 2015, 2016) play an important role for business success prediction. The latter finding is highlighted by the fact that the BPS-related variables are required for a "good" prediction accuracy for survival, high employment growth, participation in innovation subsidy programs and new patents.

Our paper unfolds as follows: we first provide some theoretical background, then present our data, subsequently introduce our empirical methods and finally discuss our results..

2. Theoretical background

Studies of the performance of firms, incumbent and new, have long been at the core of organizational research (Barney, 2001; Brush and Vanderwerf, 1992; Chandler and Hanks, 1993; Lubatkin and Shrieves, 1986; March and Sutton, 1997; Miller et al., 2013; Schendel and Hofer, 1979) and it still is a much researched topic (Delmar et al., 2003) despite growth often being termed "almost unpredictable" (Geroski, 1999, 2005) and associated with "significant uncertainty" (Catalini et al. 2019).

In fact, some scholars even claim that the growth of firms is a stochastic process, an idea first brought forward by Gibrat (1931, see Sutton, 1997). The concept of growth as a stochstic process has seen widespread use in organizational ecology (Carrol and Hannan, 2000; Hannan and Carroll, 1992; Harrison, 2004) and is also empirically tested (e.g. Wagner, 1992). More recent studies by Davis et al. (1996), Denrell (2004), Denrell et al. (2012), Geroski (1999) and Henderson et al. (2012) show that models of stochastic growth generate firm size distributions that well align with actual outcomes and interfirm differences. Such models are, however, arguably not very helpful in guiding investors, creditors and policy makers when deciding upon possible involvement in a startup. In addition, such a match between actual and predicted outcomes does not imply that performance is without cause or simply the consequence of good luck (Denrell, 2004; Denrell et al. 2012; Geroski, 1999; Henderson et al., 2012) as managers figure out how to differentiate their firms from others (Carroll, 1993) and such mangerial ability may well be reflected by the characteristics of founders and the startup itself when it enters the market.

We clearly acknowledge that chance is an important determinant of business success. However, we argue that prediction models like ours that rely on newly available, high quality "big data" may still be useful for a startups' stakeholders. In fact, randomness constitutes a first benchmark of our approach since our main prediction accuracy measure is the AUC which assumes the value is .5 if the performance outcome is generated by random chance. The AUC becomes 1 if prediction is perfect. We shall demonstrate that all our performance prediction models, even the one related to return on assets, clearly outperform random chance, indicating that stakeholders may indeed benefit from the type of performance forecasts that we suggest.

[1] See URL https://en.digst.dk/policy-and-strategy/open-government/for details.

While the firm performance literature that builts on stochastic processes is important and sizeable, the empirical literature that links alternative measures of performance outcomes and possible explanatory variables is even larger without, however, having arrived at conclusive results regarding the determinants of firm growth. Such inconclusiveness is often traced back to growth (i) being a multi-dimensional phenomenon and (ii) characterized by substantial heterogeneity (Carroll, 1993; Delmar et al., 2003). To take the multi-dimensionality problem into account, scholars have resorted to using multiple growth measures instead of focusing on a single one (Delmar et al., 2003). Carroll (1993) even argues that a single growth measure provides information about this single growth measure only. We follow Carroll's suggestion and consider five different explanatory variables, (i) involuntary exit, (ii) high employment growth, (iii) a return on assets of above 20 percent, (iv) at least one patent after foundation and (v) participation in an innovation subsidy program. These variables are, except for the last one, commonly used in management and economics. For example, Porter (1980) considers survival as a short run organizational efficiency measure and profitability as a long-run indicator and both have found widespread use. New business survival is e.g. studied by Audretsch and Mahmood (1995), Cassar (2014), Chava and Jarrow (2004), Gimmon and Levie (2010) while Morgan et al. (2009) as well as Cornett and Tehranian (1992) analyze returns on assets. Like Visintin and Pittino (2014) as well as Wennberg et al. (2011), we also consider employment growth as a main performance outcome. Patents are standard indicators for innovative activity (Blundell et al., 1995; Griliches, 1990; Kaiser et al., 2015, 2018) and particularly important to policy makers. However, not all inventions are patented and not all inventions can be patented (Arundel and Kabla, 1998). We therefore consider participation in an innovation subsidy program as an additional and broader indicator of innovative activity. All Danish innovation subsidy programs are competitive and reviewed, which in turn implies that the program sponsors assessed that the applicant firm exceeds the quality threshold for the respective subsidization program. By focusing on specific aspects of firm performance and by using distinct firm performance measures we hence take the "firm performance as a domain of separate constructs" approach in the terminology of Miller et al. (2013). We tackle the heterogeneity problem by controlling for a broad variety of variables that are found to be related to business performance and construct hitherto unexplored possible determinants, in particular those based on text-as-data. Delmar et al. (2003) argue that the conflicting results regarding performance differences between firm may be driven by a lack of accounting for elementary performance correlates like size, age, industry and governance that have already been considered by Penrose (1959) and Stinchcombe (1965). Size, age, industry and governance are variables we observe in our data which in turn means that this set of variables constitutes another relevant benchmark motivated by theory for our prediction models.

3. Data

Our core data is generated and collected by the Danish Business Authority (DBA), an administrative unit under the authority of the Danish Ministry of Business. We track all firms started between 2012 and 2014 over a period of five years. The data comprises of the universe of 55914 firms registered as limited liability companies (LLCs), joint stock corporations or a new form of a LLC called "iværksætterselskab" (IVS) whose main difference to a standard LLC is that it does not come with capital requirements and hence in effect without liabilities on part of the founders. The DBA data also provide us with the company names and addresses, NACE Rev. 2 industry codes, starting dates, total assets, profits, the number of employees as well as the names and person identifiers of their founders. In addition, the DBA data contain the BPSs since firms are obliged to report their business purpose as part of their general charters. Business purpose statements are mandatory by the Danish Law of Corporated Firms which provides firms with substantial leeway in their eventual formulation as there is no wordcount limit and the BPSs only need to loosely describe a startups' activity. As a consequence, many BPSs are very generic ("The purpose of this firm is to do trading.") while others are very specific.[2] We shall make use of this heterogeneity in our empirical analysis.

3.1. Dependent variables

We measure all our five performance variables within the first five years after establishment, except for return on assets which we measure within the first three years after foundation due to a substantial increase in missing information over a five year time horizon — many firms that started in 2014 had not yet submitted in their fifth year financial report early 2020. We define involuntary exits, and implicitly survival as well, as closures due to bankruptcy and compulsory dissolution enforced by the regulatory authorities due to non-compliance to administrative requirements. It does not include dissolution after a merger or an acquisition which would count as business success (Bates, 2005; Detienne and Wennberg, 2014; Guzman and Stern, 2015) or voluntary exits. Employment figures are provided in categories of 0, 1, 2–4, 5–9, 10–199 and more than 199 employees. We term startups that increase employment by at least two categories as "high employment growth" businesses since each category implies a doubling of the number of employees. Our final two performance measures refer to innovative activity measured as new patents and participation in an innovation subsidy program. Our patent application data originates in the "PatStat" database provided by the European Patent Office to which researchers at Copenhagen Business School have attached the unique Danish identifiers which allow us to combine our data sets (Kaiser et al., 2015, 2018). It includes all patents filed at the European Patent Office or the World Intellectual Property Organization that involve at least one

[2] E.g., "The company's purpose is to design and develop, manufacture and assemble switchboards, steering and control boards, PLC/PC/SRO solutions, automation and pre-finished projects for use by fitters, OEM/system manufacturers and the industry in general at a quality and at a price that entails that customers, suppliers and other stakeholders regard the company as an attractive and professional partner."

Danish applicant or inventor. We have data on the universe of Danish innovation support schemes collected by Danish Ministry of Higher Education and Science at our disposal. [3]

Table 1 displays a cross tabulations of high employment growth, high growth in return on assets, participation in innovation subsidy programs and new patents. Survival is left out since it is 1 by construction if any of the other outcomes is observed. The table indicates that the four dependent variables are only very tenuously related to one another, an impression formally supported by Pearson test statistics which indicate that the rows and columns in the table are indeed statistically highly significantly different for all dependent variables. Such low correlations among different performance measures is commonplace in the existing literature and it implies that different measures indeed represent different constructs (Delmar et al., 2003; Keats, 1988; Miller et al., 2013).

3.2. Explanatory variables

Our sets of explanatory variables capture proxies for the three main drivers of growth differences across firms that strategic management postulates: capabilities, management and strategy (Hopenhayn, 1992; Jovanovic, 1982; Nelson and Winter, 1982). Specifically, we relate our five performance variables to four sets of explanatory variables in addition to the "baseline" specification that will be part of all models: (i) firm name variables, (ii) BPS information, (iii) founder characteristics and (iv) startup characteristics (as well as combinations thereof). We measure all explanatory variable at the time of business foundation to account for the differences in resources on which firms initially depend (Hannan and Freeman, 1977, 1989; Scott, 1987).

Baseline variables: Our baseline specification includes measures for size, industry and governance as motivated by theory (Delmar et al., 2003; Penrose, 1959; Stinchcombe, 1965). We measure initial firm size by a dummy variable for the startup being founded by a team. Team foundations are said to have an edge over solo founders since teams pool human and financial resources instead of being dependent on the solo entrepreneur (Eesley et al., 2014; Eisenhardt and Schoonhoven, 1990), a view recently challenged by Greenberg and Mollick (2018). To account for industry affiliation, we include a set of NACE Rev. 2 one digit sector dummy variables. A missing sector classification constitutes our base category. The importance of industry has earlier been highlighted by Brüderl et al. (1992) and Clarysse et al. (2011). Our measure of governance is a dummy variable for the startup being founded by at least one other firm, e.g. as a spinoff or subsidiary. We enrich these essential theory-driven variables with information that is both easily observed in practice and that has been proven to be correlated with startup performance, namely legal form, region and founder gender. Regarding legal form and regional affiliation, the studies by Catalini et al. (2019), Guzman and Stern (2015, 2016) demonstrate that both are highly correlated with firm performance. Catalini et al. (2019, p. 18) even call legal form a prerequisite of growth. We implement legal form by dummy variables for corporations and IVSs with LLCs as our base category. Finally, Delmar and Shane (2004), Davidsson and Honig (2003) as well as Wennberg et al. (2011) show that founder gender is also related to startup performance which is why we include a dummy variable for at least one of the founders having a femake first name in our baseline specification.

(i) *Firm name variables:* Our set of firm name variables consists of information that follows Guzman and Stern (2015) and which is enriched by our own extensions. Apart from the legal form and geography variables that we consider in our baseline specificaton, the initial Guzman and Stern (2015) model contains dummy variables for (i) the firm name being eponymous (i.e. it reflects one of the founder's names; Belenzon et al. 2017), (ii) the firm name being short or long and (iii) the geographical location appearing in the firm name (specified as a dummy for any geographical location like a city, village or region appearing in the firm name and another dummy for the terms "Denmark", "Danish" or "Dan" in the firm name). A fourth component of the set of explantory variables in Guzman and Stern (2015) that we also implement in our model is a dummy variable for a startup commanding over at least one patent at the time of foundation.

We extend the initial Guzman and Stern (2015) specification by dummy variables for the firm name containing (iv) a "proper" word which we define based on the dictionary of Danish words as a proxy for the firm name containing information on what the firm actually does (like "baking", "consulting" or "plumbing), (v) the terms "holding", "capital", "invest" or "share" in the firm name to identify holding companies as well as (vi) a female name and (vii) a male name. We in addition include (viii) a founder name index since social psychology and economics suggest that person names constitute strong indicators of a persons background (Fryer and Levitt, 2004; Gerhards and Hans, 2009; Goldstein and Stecklov, 2016; Mehrabian, 1997). To account for potential information contained in founder names, we build a "name index by calculating the name-specific average performance of firms started by founders with a focal given name. We e.g. find that 85 percent of the founders named "Ulrich" survive the first five years while this is the case for 90 percent of the founders with the given name "Johan. For solo founders named Ulrich this generates an index of 0.85, for founders named Johan it is 0.9. For team foundations we take the averages across the set of founder names.

(ii) *The BPS variables:* With the use of the BPS variables we introduce a new type of information source to the literature. Similar to the use of firm name/founder name variables, we use "text-*as*-data" (Gentzkow et al., 2019). Before turning the BSP text data into explanatory variables we remove words and phrases which do not contain information relevant to our analysis, an approach called "stopping in computer linguistics. Examples for stopwords are "the", "because", "between" or "against". In addition, we "stemmed" all words in the BPSs. Stemming reduces words to their roots, e.g. the words "automation" and "automated" would

[3] This data was made available to us via the project "Investments, Incentives and the Impact of Danish Research sponsored by the Novo Nordisk Foundation.

Table 1
Crosstabulations

	RoA				Inno. subs.		
		0	1			0	1
Empl. growth	0	34421	6833	**RoA**	0	37124	719
	1	3422	930		1	7655	108
	Inno. subs.				**Patent**		
		0	1			0	1
Empl. growth	0	40669	585	**RoA**	0	37704	139
	1	4110	242		1	7753	10
	Patents				**Patent**		
		0	1			0	1
Empl. growth	0	41164	90	**Inno. subs.**	0	54103	67
	1	4293	59		1	822	86

both be reduced to their root "autom". We use the dictionary of the Danish Language Authority as our source for stemming. After stopping and stemming we define three subsets of BPS-related variables that either relate to BPS complexity, to its specificity or to its very content. As measures of BPS complexity we consider (i) the "LIX" due to Björnsson (1968) which has found widespread application in text analysis. It is calculated as the sum of the percentage of words of more than six letters and the average number of words per BPS in our context. The higher the LIX, the higher is the complexity of the text. We in addition use complexity-related variables measuring (ii) mean word length, (iii) BPS length and (iv) dummy variables for the quintiles of the BPS length distribution to put BPS length into perspective. To measure BPS specificity we use counts of how many times a "proper" word in a focal BPS appears in the universe of BPSs. We operationalize these counts as (v) the frequency with which the least common word in a focal BPS appears in the universe of BPSs and (vi) the frequency with which the most common word in a focal BPS appears in the universe of BPSs. We also control for the ratio of these two variables. Similar to our treatment of our firm name information we finally create the following content-related variables: (vii) a dummy for a geographic term appearing in the BPS, (viii) a dummy for a male name appearing in the BPS and (ix) a dummy variable for a female name appearing in the BPS. As a final subset of the BPS variables we generate (x) "wordscore indices" that measure the mean "performance" of firms' BPSs for each of our five performance indicators. The wordscore approach has been developed in political sciences where it has found widespread application in inferring political positions in text documents on the basis of scores for words derived from documents (Laver et al., 2003). It is perhaps best illustrated by providing an example. A share of 47.4 percent of the startups with the word "discotheque in their BPSs face an involuntary exit while this is true for 36.4 percent of the startups with the word "delivery in their BPSs. The wordscore associated with the word "discotheque" is defined as the words average "performance and hence is 0.474 while the other wordscore is 0.364. To aggregate the individual wordscores at the firm name level, we take the average of the individual wordscores.

(iii) *The founder characteristics:* The importance of founder characteristics and founder experience has long be recognized in entrepreneurship research (Blanchflower and Oswald, 1998; Cooper et al., 1994; Pennings et al., 1998; van Praag, 2003). We hence seek to control for these characteristics by including the following dummy variables in our predictions: (i) one of the founders having previously founded between one and three other firms and (ii) one of the founders having previously founded between more than three other firms and (iii) one of the founders having previously experienced an involuntary exit. We also include (iv) the five number of employees categories described above with this information being missing as the comparison group since firm size at startup has been shown to be highly correlated with post-entry performance (Arora and Nandkumar 2011; Bonardo et al., 2011; Brüderl et al., 1992; Clarysse et al., 2011; Delmar and Shane, 2004; Ensley and Hmieleski, 2005; Visintin and Pittini 2014; Zahra et al., 2007) and the same is true for previous founder experience (e.g. Baron and Ensley, 2006; Dencker and Gruber, 2015; Gompers et al., 2006; Westhead et al., 2005) which motivates our inclusion of the previous founding experience dummies. We additionally control for previous involuntary exit, following Cope (2011), Hayward et al. (2010), Nielsen and Sarasvathy (2016) as well as Wagner (2002).

(iv) *The startup characteristics:* Our final set of explanatory variables concerns itself with the characteristics of the startup at the time of business foundation. Financial information has long been used as a predictor for business performance (Altman, 1968, 1984; Brüderl et al., 1992; Dambolena and Khoury, 1980; Huyghebaert et al., 2000; Laitinen, 1992). We account for total assets and total profits in the first year. Both variables are missing for half of our observations, a "sparsity of data problem that is very common in big datasets. Following Gelman and Hill (2007), we set the missing explanatory variables to zero and in order to distinguish genuine 0s from the artificially created 0s introduce an additional dummy for such replacements. Since information on total assets is missing in all cases where information on profits is missing as well, we only need to include a single indicator for such a replacement having taken place. We operationalize total profits by using quantiles dummies while we take the natural logarithm of assets.

Our data contains detailed address information and we use this text-*as*-data to create indicators for the business history of each address and for the address being shared with other firms. Specifically, we include a dummy variable for at least one involuntary exit at the respective address as well as another dummy variable for nine or more involuntary exits at the address. These two dummy variables

Table 2
Descriptive statistics.

	Dummy	Mean	Std.dev.		Dummy	Mean	Std.dev.
Dependent variables				*Startup characteristics*			
Involuntary exit	1	0.172		Total assets year 1	0	12280	186841
High employment growth	1	0.095		Total assets year 1 is missing	1	0.503	
High return on assets	1	0.170		1st quintile profits year 1	1	0.099	
Innovation subsidy program	1	0.016		2nd quintile profits year 1	1	0.102	
New patent	1	0.003		3rd quintile profits year 1	1	0.092	
Explanatory variables				4th quintile profits year 1	1	0.102	
Guzman/Stern firm name variables				5th quintile profits year 1	1	0.102	
Eponymous firm name	1	0.146		Previous exit at same address	1	0.265	
Short firm name	1	0.190		>9 previous exits at same address	1	0.047	
Medium long firm name	1	0.646		Address unshared	1	0.151	
Long firm name	1	0.164		Address shared with 2–5 other firms	1	0.500	
Firm name: w/geogr. location	1	0.056		Address shared with 6–10 other firms	1	0.134	
Firm name w/Danmark, Danish, Dan	1	0.021		Address shared with >10 other firms	1	0.214	
Legal form: corporation	1	0.050		Address previously unused	1	0.132	
Legal form: IVS	1	0.099		Address previously used by 1–5 others	1	0.423	
Legal form: LLC	1	0.851		Address previously used by 6–10 others	1	0.138	
Geogr. region: Midtjylland	1	0.217		Address previously used by 11–100 others	1	0.230	
Geogr. region: Nordjylland	1	0.081		Address previously used by > 100 others	1	0.077	
Geogr. region: Sjælland	1	0.111		Sector 1	1	0.032	
Geogr. region: Syddanmark	1	0.165		Sector 2	1	0.145	
Geogr. region: Greater Copenhagen	1	0.425		Sector 3	1	0.039	
Patents at foundation	1	0.004		Sector 4	1	0.161	
Extended Guzman/Stern firm name variables				Sector 5	1	0.062	
In firm name: a proper danish word	1	0.475		Sector 6	1	0.029	
In firm name: Holding, capital, shares	1	0.282		Sector 7	1	0.012	
In firm name: female name	1	0.081		Sector is missing	1	0.520	
In firm name: male name	1	0.158		Mean ind. perf. invol. exit	0	0.173	0.084
Founder name index invol. exit	0	0.171	0.070	Mean ind. perf. high empl. growth	0	0.079	0.057
Founder name high empl. growth	0	0.082	0.035	Mean ind. perf. high ret. on assets	0	0.142	0.057
Founder name high ret. on assets	0	0.135	0.038	Mean ind. perf. new patents	0	0.016	0.038
Founder name new patents	0	0.018	0.012	Mean ind. perf. innov. subsidy program	0	0.005	0.024
Founder name innov. subsidy program	0	0.006	0.006	*BPS information*			
Human capital variables				LIX	0	54.0	9.6
At least one founders is firm	1	0.369		Mean word length	0	9.4	2.2
At least one of founder has female first name	1	0.180		BPS lengths	0	41.3	33.0
At least one of founder has male first name	1	0.868		BPS 1st quintile	1	0.192	
Team	1	0.111		BPS 2nd quintile	1	0.195	
# employees year 1: 0	1	0.042		BPS 3rd quintile	1	0.201	
# employees year 1: 1	1	0.068		BPS 4th quintile	1	0.203	
# employees year 1: (2,4)	1	0.069		BPS 5th quintile	1	0.208	
# employees year 1: (5,9)	1	0.030		Frequency of least common word	0	1247	2160
# employees year 1: (6,49)	1	0.025		Frequency of most common word	0	5673	3818
# employees year 1: missing	1	0.767		Freq. least/freq. most common word	0	0.291	0.370
No previous founding experience	1	0.198		A geogr. term name is in BPS	1	0.020	
Previously founded 1–3 firms	1	0.463		A male name is in BPS	1	0.022	
Previously founded more than 3 firms	1	0.212		A female name is in BPS	1	0.011	
Earlier invol. exit by one founder	1	0.084		BPS wordscore invol. exit	0	0.163	0.071
				BPS wordscore high empl. growth	0	0.080	0.045
				BPS wordscore high ret. on assets	0	0.144	0.045
				BPS wordscore new patents	0	0.019	0.020
				BPS wordscore innov. subsidy program	0	0.006	0.009

Standard deviations are displayed for continuous variables only.

may serve as proxies for the overall attractiveness of the location and other characteristics associated with a given address. We control for how many other firms reside under the same address since many corporations often co-reside with their associated holding companies by including dummies for the address being shared by 2–5, 6–10 and more than 10 other firms with the address being unshared being the base category. In addition, we account for the present address having previously been used by 1–5, 6–10, 11–100 and more than 100 firms.

To more precisely account for industry heterogeneity than with just the set of sector dummy variables used in our baseline specification without being forced to include a large set of dummy variables, we include mean industry performance for all our five performance indicators which we calculate on the basis of the Danish Industry Classification that is slightly more detailed than NACE Rev. 2 four digit level.[4]

[4] For example, if 30 percent of *other* firms in the regression sample in a focal firms industry experience high employment growth, the associated mean performance for in this industry is 0.3.

Table 2 presents descriptive statistics of our dependent and explanatory variables. It shows that involuntary exits are comparatively rare events with 17.2 percent of the firms in our data involuntarily exiting within five years of operation, a figure that is substantially lower than the 50 percent overall exits reported by e.g. Headd (2003) or Mata and Portugal (1994). Different to those studies we focus on firms with a legal form that requires registration and consider the universe of startups instead of merely technology-driven ones. A tenth of the firms in our data generate substantial employment growth while 17 percent achieve a high return on assets. By contrast, participation in an innovation subsidy program and taking out a new patent are both rare events. A mere 1.6 percent of our firms participate in innovation subsidy programs while only 0.3 percent apply for a new patent within their first five years of existence.

More than 40 percent of all startups are founded in the capital greater Copenhagen region, only 0.4 percent of all firms has applied for a patent at the time of foundation, about a third of the startups involve another firm as a founder, 87 percent of the startups are founded by men, more than 89 percent are solo foundations and 46 percent are founded by serial entrepreneurs which compares to a European average of 30 percent and a US average of 13 percent (Plehn-Dujowich, 2010). Turning to the information contained in the BPSs, the average LIX is 54 which is considered as "difficult" by Björnsson (1968). The mean word length is 9.4 characters while average BSP lengths is 41.3 characters.

The correlations between our explanatory variables are modest with our largest variance inflation factor being 2.56 which is well below the critical value of 10 (Belsley et al., 1980). This is reinforced by Appendix A which displays the correlation coefficients.

4. Empirical analysis

4.1. Empirical strategy

Our empirical aim is twofold: we want to analyze (i) the degree of accuracy to which publicly available data can be used to forecast business startup performance and (ii) what sets of variables — and combinations thereof — are best at predicting performance since not all variables may be equally easy to get a handle on. We seek to achieve our goals by subsequently introducing the (i) firm name variables, (ii) BPS variables, (iii) founder characteristics and (iv) startup characteristics (and their combinations) to our baseline set of performance predictors. The baseline set of variables hence constitutes our main benchmark. We run binary logit performance regressions and subsequently assess the out-of-sample prediction accuracy of our specifications. We estimate our models on a 70 percent random sample and retain the remaining 30 percent for prediction, following Guzman and Stern (2015). We calculate our firm name indices and our BPS wordscores as well as the average industry performance index on the regression sample and extrapolate them to our holdout sample.

Our focus is on the prediction of outcomes which is why we present the forecasting accuracy statistics only and relegate logit coefficient estimation results for our full models to Appendix B. We apply one main prediction accuracy measure, the AUC and also briefly discuss the pseudo R^2due to McFadden (1973). Our focus is on the AUC as a standard measure of forecast performance of binary models (Hand, 2001). It illustrates the performance of a classification model like ours by plotting the observed rate of outcomes against the rate of false positive outcomes at pre-specified threshold levels (the receiver-operator curve, ROC) — deciles in our case as in Cooper (1993). The area under thes curve is a measure of predictive accuracy. Bradley (1997) defines a model that corresponds to an AUC of between 0.5 and 0.6 as a "fail', values between 0.6 and 0.7 as "poor", between 0.7 and 0.8 as "fair", between 0.8 and 0.9 as "good" and values above 0.9 as "excellent". We also calculate the change in AUC relative to the baseline model. We use the pseudo R^2 as an additional measure of prediction accuracy since it is an often used goodness of fit statistic and is reported by all standard software packages by default. Neither the AUC nor the pseudo R^2 penalize the degrees of freedom which is, however, irrelevant given the size of our data any other administrative dataset.

In addition to the AUC and the pseudo R^2, we display the true positive rate (TPR, the ratio of all correctly predicted positive outcomes relative to all positive outcomes; also termed "specificity", "recall", "precision" or "hit rate") and the true negative rate (TNR or "selectivity" and "specificity").[5] Specificty and selectivity are widely discussed in even the broader public in the context of tests for COVID19 infections and antibodies and are hence more easily interpretable than the AUC or the pseudo R^2.

All our models include the baseline set of explanatory set of variables which allows us to compare the log-likelihood basic model to the richer models (Greene, 2017; Wooldridge, 2010). These test statistics cannot reject that all models that include variables beyond the baseline ones are jointly statistically highly significant and that adding additional sets of variables adds explanatory power. We therefore do not display the corresponding test statistics or the associated *p*-values in our results table.

4.2. Results

Table 3 presents our prediction outcomes, e.g the AUC and the pseudo R^2 along with specificity and selectivity. This subsection focuses on the AUC as an aggregate measure of TNR and TPR as well as the pseudo R^2. Since TNR and TPR are arguably more accessible to practitioners we shall discuss them in our discussion section, Section 5.

A first striking finding is that the information contained in our BPS-related variables is rich enough to *alone* predict the two innovation–related outcomes with "good" accuracy. An even "excellent" accuracy is achieved for participation in an innovatrion subsidy

[5] The two other relevant quantities are the false negative rate (FNR or "miss rate") with FNR = 1-TPR and the false positive rate (FPR or "fall-out") with FPR = 1-TNR.

program once the BPS data is combined with the firm name variables. An "excellent" predictive performance is not obtained for any other performance outcome. A second striking result is that all our specifications poorly predict a high return on assets. Even though adding additional sets of explanatory variables, and here in particular the BPS-variables, leads to a massive improvement in predictive accuracy as measured by the AUC, it still remains "poor" with a maximum AUC of 0.686.

Business survival is predicted with "good" accuracy and an AUC of 0.801 once the two sets of text data is combined with the set of founder characteristics. Predictive power can be increased by 2.7 percentage points if the BPS variables are added as well, leading to a maximum AUC of 0.823. Adding even more variable sets does, however, not increase predictive power. Similarly, it also needs the combination of at least two additional sets of variables aprt from the baseline ones, the BPS variables and the set of founder charactertistics, to attain "good" predictive accuracy for high employment growth.

Turning to the pseudo R^2s as alternative prediction accuracy measures, we find that the AUC and the pseudo R^2 do not always point at the same preferred specification. This is a consquence of the pseudo R^2 being based on the maximum likelihood value while the AUC is calculated on the basis of the distribution of predicted vs. actual outcomes. However, there is a very high correlation between both statistics with a correlation coefficient of 0.88.

4.3. Robustness checks

Even though all data we use in our analysis is publicly available, not all variables may be easily obtainable in all countries. In addition, not all variables are equally simple to generate and to process. In our robustness checks we therefore test the extend to which leaving out such type of information affects prediction accuracy. We in particular reckon that the initial number of employees, the initial number of patents and the initial financial variables may not be easily obtainable in all countries and that the BPS wordscore index, the mean industry performance index and the founder name index may not easily be computed for industry practitioners.

Appendix C displays the AUCs of more restricted specifications where we leave out initial firm size, initial patents, the initial financial variables, the founder names index, the BPS wordscores and mean industry performance. It shows that the strongest decreases in prediction accuracy go along with leaving out the BPS wordscore index and/or the mean industry performance index while leaving out the founder name index has little effect on forecasting performance. Omitting the BPS wordscores decreases AUC by 2.8 percentage points for survival and by 2.2 percentage points for high returns on assets. It does, however, not lead to a reclassification of prediction accuracy in either case. A reclassification of prediction accuracy from "good" to "fair" is encountered for high employment growth if the industry performance index is left out. Similiarly, prediction accuracy drops from "excellent" to "good" for participation in an innovation subsidy program if this index is omitted. The industry performance index overall leads to redcutions of 1.9, 5.2, 3.9. 6.6 and 5.1 percentage points for each of our five respective performance indicators. These changes are also reflected in the change in prediction accuracy due to leaving out all three indices. If both the BPS wordscore, the mean industry performance index and the founder name index are not part of the specification, predictive accuracy drops by between 2.4 and 6.6 percentage points for our five variables of interest.

By contrast, leaving out the founder name index, the financial variables and initial patents has very little effect on prediction accuracy. Omitting initial firm size only substantially reduces the predictive accuracy for high employment growth which decreases by 3.5 percentage points compared to the full model, indicating state dependence in firm size (Audretsch et al., 1999; Geroski, 1999).

We hence conclude that leaving out variables like the initial number employees, initial patents and initial financial information that may not be easily accessible has little effect on overall predictive power. This is different to the set of indices that we use which is likely to be available but which is less conveniently computed. While the founder name index plays little role in forecasting accuracy, the BPS wordscores and in particular the mean industry performance index constitute important determinants of predictive power. However, generating the mean industry performance index essentially entails taking averages across industries and hence may not be prohibitively complicated to compute.

5. Discussion

5.1. Overview

Our models predict startup survival, high employment growth and new patents well. They predict participation in an innovation subsidy program even very well but fail to predict high returns on assets with acceptable accuracy. We have shown that it is sufficient to include the BPS-related text-*as*-data variables alone to generate a "good" predictive accuracy for new patents and participation in an innovation subsidy program (in combination with our "baseline" variables legal form, region, ownership and industry classification). To get an "excellent" predictive accuracy for participation in an innovation subsidy program the BPS variables need to be combined with the firm name variables, another set of text-*as*-data information. These two sets of text-*as*-data are also required to generate a "good" prediction for startup survival when combined with the set of initial founder characteristics. To achieve a "good" prediction of high employment growth, a combination of the "basic" variables the BPS-derived variables and the founder characteristics is required.

We hence demonstrate that it is possible to predict startup performance as measured by survival, high employment growth, new patents and participation in an innovation subsidy program with a considerable degree of accuracy based on publicly available real-time data alone using a particularly simple econometric approach, the binary logit model. The combination of publicly available data and the simplicity of the econometric analysis deems our approach very accessible to practitioners and academics alike. Below, we summarize the implications our study has for theory and practice, discuss its limitations and finally conclude.

Table 3
Prediction accuracy.

Sets of variables included					ROC-AUC			Pseu-do	Sensitivity (TPR)		Specificity (TNR)	
Firm name	BPS	Founder char.	Startup chars	dof	Val.	Cat.	Δ	R^2	Val.	Δ	Val.	Δ
Survival												
				15	0.664	poor		0.055	0.737		0.519	
x				28	0.700	fair	5.5	0.099	0.685	7.1	0.613	18.1
	x			29	0.746	fair	12.3	0.117	0.729	1.1	0.639	23.0
		x		24	0.750	fair	13.0	0.128	0.721	2.2	0.664	27.9
			x	31	0.745	fair	12.2	0.135	0.651	11.6	0.698	34.5
x	x			42	0.752	fair	13.2	0.143	0.727	1.5	0.641	23.5
x		x		37	0.770	fair	16.0	0.164	0.715	3.1	0.706	35.9
x			x	44	0.757	fair	14.0	0.166	0.691	6.3	0.690	33.0
	x	x		38	0.796	fair	19.8	0.183	0.742	0.7	0.706	35.9
	x		x	45	0.784	fair	18.1	0.175	0.727	1.4	0.699	34.7
		x	x	40	0.799	fair	20.3	0.203	0.703	4.7	0.740	42.5
x	x	x		51	0.801	good	20.6	0.205	0.747	1.3	0.706	35.9
x	x		x	58	0.785	fair	18.2	0.194	0.731	0.8	0.700	34.8
x		x	x	53	0.806	good	21.3	0.227	0.736	0.2	0.725	39.7
	x	x	x	54	0.822	good	23.9	0.236	0.751	1.9	0.733	41.3
x	x	x	x	67	0.824	good	24.1	0.252	0.753	2.1	0.732	40.9
High employment growth												
				15	0.695	poor		0.053	0.686		0.617	
x				28	0.719	fair	3.5	0.112	0.807	17.6	0.526	−14.8
	x			29	0.794	fair	14.2	0.144	0.797	16.2	0.649	5.1
		x		24	0.735	fair	5.9	0.095	0.684	0.2	0.684	10.8
			x	31	0.702	fair	1.0	0.101	0.620	9.6	0.675	9.4
x	x			42	0.788	fair	13.5	0.167	0.809	18.0	0.629	1.8
x		x		37	0.769	fair	10.8	0.157	0.774	12.8	0.622	0.8
x			x	44	0.731	fair	5.2	0.154	0.731	6.6	0.608	−1.5
	x	x		38	0.829	good	19.4	0.195	0.779	13.6	0.724	17.2
	x		x	45	0.789	fair	13.6	0.178	0.730	6.4	0.693	12.3
		x	x	40	0.739	fair	6.3	0.142	0.637	7.1	0.724	17.3
x	x	x		51	0.826	good	18.9	0.220	0.799	16.5	0.704	14.1
x	x		x	58	0.786	fair	13.1	0.198	0.759	10.6	0.674	9.1
x		x	x	53	0.771	fair	11.1	0.196	0.723	5.4	0.674	9.2
	x	x	x	54	0.823	good	18.5	0.225	0.736	7.3	0.740	19.8
x	x	x	x	67	0.822	good	18.3	0.249	0.759	10.6	0.723	17.2
High return on assets												
				15	0.585	fail		0.014	0.593		0.539	
x				28	0.599	fail	2.2	0.033	0.561	−5.5	0.592	10.0
	x			29	0.661	poor	13.0	0.049	0.656	10.5	0.599	11.2
		x		24	0.610	poor	4.2	0.029	0.681	14.8	0.471	−12.6
			x	31	0.637	poor	8.7	0.059	0.607	2.3	0.599	11.2
x	x			42	0.652	poor	11.3	0.061	0.627	5.7	0.606	12.5
x		x		37	0.622	poor	6.3	0.048	0.606	2.2	0.571	6.0
x			x	44	0.638	poor	8.9	0.074	0.580	−2.2	0.629	16.8
	x	x		38	0.674	poor	15.1	0.063	0.681	14.8	0.578	7.3
	x		x	45	0.676	poor	15.4	0.080	0.644	8.5	0.617	14.5
		x	x	40	0.652	poor	11.4	0.069	0.651	9.7	0.573	6.3
x	x	x		51	0.665	poor	13.7	0.074	0.652	9.8	0.593	10.2
x	x		x	58	0.668	poor	14.1	0.091	0.629	6.0	0.626	16.2
x		x	x	53	0.652	poor	11.4	0.084	0.614	3.4	0.611	13.4
	x	x	x	54	0.686	poor	17.1	0.091	0.666	12.2	0.601	11.6
x	x	x	x	67	0.678	poor	15.9	0.101	0.644	8.6	0.618	14.7
Innovation subsidy program												
				15	0.758	fair		0.091	0.708		0.636	
x				28	0.820	good	8.2	0.167	0.792	12.0	0.675	6.2
	x			29	0.891	good	17.5	0.274	0.815	15.2	0.810	27.5
		x		24	0.786	fair	3.7	0.112	0.692	−2.2	0.718	12.9
			x	31	0.787	fair	3.7	0.180	0.673	−4.9	0.750	18.0
x	x			42	0.901	excellent	18.8	0.314	0.846	19.6	0.803	26.3
x		x		37	0.832	good	9.7	0.184	0.773	9.2	0.702	10.5
x			x	44	0.832	good	9.7	0.235	0.765	8.2	0.723	13.7
	x	x		38	0.900	excellent	18.7	0.296	0.804	13.6	0.815	28.2
	x		x	45	0.891	good	17.5	0.307	0.788	11.4	0.823	29.5
		x	x	40	0.802	good	5.8	0.197	0.681	−3.8	0.762	19.9
x	x	x		51	0.908	excellent	19.8	0.333	0.850	20.1	0.814	28.0
x	x		x	58	0.899	good	18.6	0.343	0.823	16.3	0.814	28.1

(*continued on next page*)

Table 3 (*continued*)

Sets of variables included					ROC-AUC			Pseu-	Sensitivity		Specificity	
Firm		Founder	Startup					do	(TPR)		(TNR)	
name	**BPS**	**char.**	**chars**	**dof**	**Val.**	**Cat.**	**Δ**	R^2	**Val.**	**Δ**	**Val.**	**Δ**
x		x	x	53	0.839	good	10.6	0.248	0.746	5.4	0.743	17.0
	x	x	x	54	0.896	good	18.1	0.324	0.792	12.0	0.828	30.2
x	x	x	x	67	0.904	excellent	19.2	0.358	0.819	15.8	0.824	29.7
New patent												
				15	0.703	fair		0.130	0.520		0.763	
x				28	0.791	fair	12.5	0.340	0.560	7.7	0.864	13.3
	x			27	0.853	good	21.4	0.281	0.640	23.1	0.853	11.8
		x		24	0.704	fair	0.1	0.148	0.480	−7.7	0.795	4.1
			x	31	0.716	fair	1.9	0.219	0.500	−3.8	0.797	4.5
x	x			40	0.860	good	22.3	0.398	0.660	26.9	0.886	16.1
x		x		37	0.797	fair	13.3	0.355	0.620	19.2	0.875	14.7
x			x	44	0.777	fair	10.5	0.367	0.520	0.0	0.865	13.4
	x	x		36	0.853	good	21.3	0.294	0.640	23.1	0.860	12.8
	x		x	43	0.837	good	19.0	0.326	0.600	15.4	0.865	13.3
		x	x	40	0.725	fair	3.2	0.235	0.500	−3.8	0.818	7.1
x	x	x		49	0.863	good	22.8	0.413	0.680	30.8	0.894	17.2
x	x		x	56	0.832	good	18.4	0.422	0.640	23.1	0.890	16.6
x		x	x	53	0.790	fair	12.3	0.381	0.580	11.5	0.877	15.0
	x	x	x	52	0.838	good	19.2	0.338	0.620	19.2	0.870	14.0
x	x	x	x	65	0.840	good	19.6	0.434	0.640	23.1	0.896	17.5

The "base" set of variables is included in all specifications. Δ refer to the percentage change in the corresponding value ("Val.") relative to the baseline model. "dof" denotes the degrees of freedom of the respective estimation model.

5.2. Theoretical implications

This paper essentially constitutes a measurement exercise: we put ourselves in the shoes of important stakeholders in startups in an attempt to help them process publicly available, real-time information in an effective way to improve their performance forecasts. Investors, banks and policy makers are interested in the survival of a focal startup. Banks are additionally concerned about default, e.g. survival, while investors expect high growth in terms of employment and return on assets, conditional on business survival (Guzman et al. 2019). Designers of innovation policies will be most interested in the innovative performance of the firms that may qualify for their subsidy programs.

Despite our work primarily aiming at better informing stakeholders, it bears a number of theoretical implications. First, while we do in fact predict desirable performance outcomes with "good" accuracy, our predictions are far away from being perfect. This in turn may imply that valued outcomes may be less predictable than what managers believe" as Denrell et al. (2015, p. 936) put it. Such randomness requires managers and stakeholders to remain being flexible and that randomness needs to be taken seriously indeed (Denrell et al., 2015; Mintzberg, 1990). Second, we show that text-*as*-data plays an important role in achieving accurate performance predictions. While empiricists have embraced such data for quite some time (Gentzkow et al., 2019), theory has yet been little concerned with the actual meaning and the implications of such new types of data. For example, we show that text contained in BPSs contains valuable information for startup prediction but we do not yet know why entrepreneurs phrase their purpose statements the way they do and what the implications of research like ours has on the formulation of future BPSs. Third, much existing empirical research on startup performance has grappled with highly selective data (Parker, 2008), thereby possibly wrongly informing theory about the true underlying empirical mechanisms. Publicly available data on entire populations of startups will hence greatly improve the way empirics informs theory.

5.3. Practical implications

The key practical implication of our paper of course is that we show that it indeed is possible to forecast startup performance using publicly available data and simple econometrics. Our key prediction accuracy measure, the receiver-operator area under the curve, does, however, not have an easily accessible interpretation apart from a value of .5 referring to tossing a coin and a value of 1 representing perfect prediction. As a first step of our discussion of practical implications we therefore discuss sensitivity (the true positive rate TPR) and specificity (the true negative rate) with the "false positive rate" (FPR or "type I error") and the "false negative rate" (FNR or "type II error") given by FPR = 1-TNR and FNR = 1-TPR. As an aggregate measure of sensitivity and specificty, the highest values of the AUC do not necessarily coincide with the highest value of sensitivity or specificity; indeed, the AUC trades sensitivity against specificity such that similar model performance rankings based on either AUC or specificty/sensitivity are only achieved for similar values of sensitivity and specificity.

There is a big difference in specificity and sensitivity from a practitioner's persepective: a high sensitivity implies "picking winners" while a high specificity implies "avoiding losers". Practitioners may argue that wrongly predicting winners may be more costly than wrongly predicting losers since wrongly predicting winners entails lost financial engagenment while wrongly predicting losers entails

lost opportunities. The cost of misclassification are hence not equal in real world settings. Specificity may hence be preferred from a real life perspective relative to selectivity (Gepp et al., 2010; Wang et al., 2014).

Our best performing models in terms of specificity for survival and high employment growth yield TNRs of around 72 percent while we get specificity values of 82 percent for participation in innovation subsidy programs and of 90 percent for new patents. The best prediction for high returns yields a specificity of 63 percent only. For all our performance measures except for new patents we find that specificty and sensitivity are roughly similar which in turn implies that the model performance rankings obtained using AUC are similar to those obtained by either specificity or sensitivity. The exception is new patents where sensitivity is about 20 percentage points lower than specificity. For new patents it is hence easier to avoid loosers than to pick winners.

Apart from our model actually being able to forecast startup performance with considerable accuracy another important practical implication is that test-*as*-data may constitute a major pillar in business performance predictions. Guzman and Stern (2015, 2016), Catalini et al. (2019) have already demonstrated that there is valuable information in firm names and in the mapping between firm name and founder names. While we also find some predictive power in firm names, we show that there is much more predictive power in business purpose statements. Stakeholders should hence scrutinize such text data to a similar extend as financial variables and the CVs of the founders, in particular since we find that both financial variables and founder characteristics have lower predictive power than the BPS-related variables. Scholars have already demonstrated that there is predictive power in the texts related to stock listed firms (Antweiler and Frank, 2004; Tetlock, 2007) and it may be now time to also use text-*as*-data for possible involvements in startups.

Given that many government agencies are increasingly smart about using data analysis to improve their operations and services as stated by Einav and Levin (2013, p. 12), we believe that the allocation decisions of government funds will increasingly often be backed by performance prediction models similar to ours. It is a point in case that we were made aware of the existence of the BPS data by a Danish governmental agency.

5.4. Limitations and future directions

There are several "constraints on generality" (Simons et al., 2017) that apply to our paper. First, we have used data on one single country only. It is of course not clear if our results carry over to other countries as well. However, given the massively increased availability of data similar to ours across the world, we do believe it will soon be possible to replicate our study on a broader scale. Moreover, Uhlbach et al. (2019) show that entrepreneurship rates in Denmark are not much different from from those in Sweden or the UK. In addition, Denmark offers particularly business-friendly environment, ranking third behind New Zealand and Singapore and ahead of the US and the UK. [6] While this may indicate that the Danish entrepreneurial ladscape is perhaps not too different from larger economies, a boundary condition for our work to be generalizable clearly is the availability of high quality public and population-wide data. In addition, our estimates concern a period of distinct growth in the Danish economy and it is of course possible that prediction accuracy varies across different phases of the business cycle, let alone exogeneous shocks like a pandemic or other natural disasters.

Following up on the important data aspect of our paper, assembling our data set in the degree of completeness it now has achieved has been a major effort. To arrive at accurate performance predictions we would, however, not have needed a data set as comprehensive. As discussed above, a major pillar of our forecasts is text-*as*-data such as variables derived from firm names and the business purpose statements. Arguably harder to get information like data on initial patents, initial firm size and initial financial indicators by contrast add relatively little to forecasting performance. If one wants to economize on data collection efforts her focus should be on the textual information.

Another limitation of our work is that while we do account for heterogeneity among the startups we consider by our large set of explanatory variables, we assume that the coefficient estimates we generate are identical across all firms. However, as Delmar et al. (2003, p. 190) find that, "growth can be achieved in a number of different ways, and the pattern of firm growth, over time, can look very different across all growth firms". This implies that we most likely could further improve prediction accuracy by splitting up our data e.g. according to industry as suggested by Harrison (2004) who in fact calls for separate growth models for separate industries.

Throughout this paper we have solely relied upon simple binary logit models even though machine learning methods have become increasingly popular to study investment decisions (Ghassemi et al., 2020; Krishna et al., 2016). Apart from these models also being much more complicated and less accessible to practitioners in particular, they may also not lead to improvements in prediction accuracy. Krishna et al. (2016) for e.g. show that the simple logit model works as well as a set of standard machine learning models they apply in addition. We doubt that machine learning methods would improve forecasting performance in our context since these approaches are bound to only improve prediction accuracy in cases where the number of possible explanatory variables is large but the number of observations is small (Taddy, 2013) — which is not the case in our setting. In fact, using a LASSO model, a frequently applied machine learning technique, instead of our simple logit model generates prediction accuracies that are almost exactly the same. More importantly perhaps, the number of variables eventually selected by the LASSO models is almost exactly the same as the number of variables used in our best performing logit models, implying that the LASSO models neither improve predictive accuracy nor save on the associated degrees of freedom. However, machine learning models like LASSO may indeed be preferable to simple logit models if our data was split up in order to better account for the heterogeneity of startups since this implies a reduction in the number of observations.

Future work may also consider additional dependent variables like IPO events that are studied by Guzman and Stern (2015, 2016) as well as Catalini et al. (2019). IPOs may, however, be arguably more relevant for Silicon Valley than for Denmark as the Crunchbase database of May 23, 2020 only lists 18 investment events of Danish firms founded a year ago and not a single IPO. Still, the database lists 561 investment over the time horizon we consider, deeming a deeper look at such events a viable alternative performance indicator.

[6] World Bank/IBRD: Doing Business 2018 Reforming to Create Jobs.

Finally, we may consider using our work for "nowcasting" (Banbura et al., 2013; Gentzkow et al., 2019), i.e. to use our combination of different data sources to predict our key performance measures in real-time speed which is possible since the data we use is provided on a real time, e.g. daily, basis. This may eventually lead to algorithmic scoring models that can possibly even work in real time, similar to the risk scores for health care problems discussed in Einav and Levin (2013).

6. Conclusions

Easily accessible and publicly available data, both textual and non-textual, are starting to become accessible in most modern economies. We show how such data can be used to predict the expected performance of newly started enterprises with substantial accuracy. Such performance predictions are of great importance to investors, creditors and policy makers alike. Investors may not only want to assess the prospects of a business that asks for funding, they may also be interested in identifying promising startups before they even apply. Some investors have already embraced "algorithmic scoring" models (Corea, 2018; Diffey, 2019; Palmer, 2017) and our paper indicates that it is well possible to successfully implement such methods. Even though banks are unlikely to be equally proactive, they may as well want to more firmly base their debt financing decisions on objective data-driven grounds. Lastly, policy makers may gain from the improved identification of promising startups in order to be better gear innovation support programs towards such firms and to improve the tailoring of startup promotion programs more generally.

For our predictions, we use data on the universe of Danish firms started between 2012 and 2014 to run simple logit regressions to show that key performance outcomes such as survival, employment growth, patenting activity as well as participation in competitive and audited innovation support programs can be predicted with high accuracy using publicly available data alone. Our models essentially only require text-as-data information that startups have to report when they register: startup names, founder identities, addresses and business purpose statements. Even though including hard-to-get or hard-to-process additional information on initial firm size, initial patents and an wordscore index constructed from the buiness purpose statements improves prediction accuracy, such more intricate data is not necessary to forecast startup performance with substantial precision. However, even our most complex model was unable to predict returns on asset of above 20 percent with even modest accuracy.

Given that may countries worldwide have opened or will open their data treasures worldwide (Gentzkow et al., 2019) and that the data we use have recently become publicly available through the open data policy adopted by the Danish government in 2011 as part of the global "Open Government Partnership" initiative highlights that such open data policies indeed are effective in improving economic decision making at very low cost.

Appendix A. Table of correlations

		1	2	3	4	5	6	7	8	9	10	11	12	13
1	Legal form: corporation	1.00												
2	Legal form: IVS	−0.08	1.00											
3	Geogr. region: Midtjylland	0.02	−0.02	1.00										
4	Geogr. region: Nordjylland	0.01	0.00	−0.16	1.00									
5	Geogr. region: Sjælland	−0.03	−0.01	−0.19	−0.11	1.00								
6	Geogr. region: Syddanmark	0.03	−0.03	−0.23	−0.13	−0.16	1.00							
7	One founder is firm	0.19	−0.14	0.02	−0.01	−0.04	−0.01	1.00						
8	One of founder w/ female first name	−0.08	0.05	−0.02	−0.01	0.04	−0.01	−0.30	1.00					
9	Team	−0.05	0.08	0.00	0.00	0.00	0.00	−0.20	0.31	1.00				
10	Sector 1	0.07	−0.04	0.02	0.03	0.01	0.03	0.05	−0.02	0.01	1.00			
11	Sector 2	0.02	−0.07	0.01	0.01	0.04	0.01	0.08	0.00	0.00	−0.07	1.00		
12	Sector 3	−0.02	−0.04	−0.03	−0.01	0.00	−0.01	0.04	0.02	0.02	−0.04	−0.08	1.00	
13	Sector 4	−0.01	−0.10	0.01	−0.03	0.00	−0.01	−0.01	−0.01	−0.02	−0.08	−0.18	−0.09	1.00
14	Sector 5	0.00	−0.02	−0.01	−0.01	−0.01	−0.03	0.06	−0.01	0.00	−0.05	−0.11	−0.05	−0.11
15	Sector is missing	−0.02	0.17	0.01	0.02	−0.04	0.01	−0.11	−0.01	0.00	−0.19	−0.43	−0.21	−0.46
16	In firm name: a proper danish word	0.03	−0.05	0.00	0.02	0.02	0.02	0.07	0.00	−0.02	0.01	0.07	0.04	−0.05
17	In firm name: Holding, capital, shares	−0.09	−0.05	0.04	0.03	−0.01	0.03	−0.34	0.04	−0.02	−0.11	−0.25	−0.12	0.16
18	In firm name: female name	−0.01	−0.01	0.00	0.00	−0.01	0.00	−0.03	0.13	−0.01	−0.01	−0.02	0.03	−0.01
19	In firm name: male name	−0.03	−0.02	0.01	0.01	0.01	0.01	−0.09	−0.02	−0.04	−0.02	0.00	0.00	−0.01
20	Eponymous firm name	−0.07	0.03	0.02	0.01	0.01	0.02	−0.31	0.09	−0.02	−0.04	−0.06	−0.05	0.07
21	Short firm name	−0.01	0.04	−0.01	−0.01	0.00	−0.01	0.02	0.00	0.02	0.02	0.03	0.02	−0.02

(continued on next page)

(*continued*)

		1	2	3	4	5	6	7	8	9	10	11	12	13
22	Long firm name	0.02	−0.04	0.02	0.02	0.01	0.02	−0.01	−0.02	−0.02	−0.01	−0.03	−0.02	0.01
23	Firm name: w/geogr. location	0.04	−0.04	0.03	0.06	0.01	0.05	0.05	−0.02	−0.01	−0.01	0.02	0.01	−0.02
24	Firm name w/ Danmark, Danish, Dan	0.05	0.00	−0.01	0.00	0.00	0.01	0.03	−0.01	0.00	0.02	0.03	−0.01	−0.02
25	LIX	0.06	−0.01	−0.01	0.02	−0.02	−0.01	0.02	−0.03	−0.01	0.02	−0.10	−0.04	0.03
26	Mean word length	−0.01	−0.05	0.00	0.01	0.00	0.01	−0.09	0.01	−0.02	−0.02	−0.13	−0.01	0.06
27	BPS lengths	0.09	−0.03	−0.01	0.01	−0.03	−0.02	0.06	−0.04	0.00	0.02	−0.09	−0.04	0.03
28	BPS 2nd quintile	−0.04	0.00	0.00	−0.01	0.02	0.00	−0.02	0.02	0.00	−0.01	0.02	0.01	0.01
29	BPS 3rd quintile	−0.02	−0.01	0.01	−0.01	0.00	0.02	0.00	−0.01	−0.01	0.00	0.01	−0.01	0.01
30	BPS 4th quintile	0.00	−0.02	0.02	0.00	−0.01	0.01	0.00	−0.02	0.00	0.00	−0.03	−0.02	0.04
31	BPS 5th quintile	0.07	−0.01	−0.02	0.03	−0.02	−0.02	0.04	−0.02	0.00	0.02	−0.08	−0.03	0.01
32	Frequency of least common word	−0.09	−0.01	0.03	0.01	0.01	0.01	−0.15	0.00	−0.04	−0.11	−0.09	−0.06	0.10
33	Frequency of most common word	−0.01	−0.01	0.04	0.02	−0.03	0.02	−0.06	−0.02	−0.03	−0.05	−0.07	−0.11	0.10
34	Freq. least/freq. most common word	−0.05	0.00	−0.01	−0.01	0.02	0.01	−0.09	0.03	−0.01	−0.05	−0.02	0.04	0.00
35	A geogr. term name is in BPS	0.05	−0.02	0.04	−0.01	−0.02	0.00	0.07	−0.02	−0.01	0.00	−0.02	0.00	0.02
36	A male name is in BPS	0.01	0.00	0.00	0.01	0.01	0.00	0.01	0.00	0.01	0.01	0.03	0.02	−0.03
37	A female name is in BPS	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.00
38	# employees year 1: 0	0.01	−0.01	−0.01	0.00	0.01	0.00	0.02	0.01	0.00	0.03	0.06	0.05	−0.03
39	# employees year 1: 1	−0.01	−0.02	0.01	0.00	0.01	0.01	0.02	−0.01	−0.04	0.04	0.10	0.00	−0.03
40	# employees year 1: (2,4)	0.03	−0.05	0.00	0.01	0.02	0.01	0.04	0.02	0.03	0.03	0.13	0.04	−0.06
41	# employees year 1: (5,9)	0.07	−0.05	0.00	0.00	0.02	0.01	0.05	0.00	0.00	0.03	0.07	0.07	−0.05
42	# employees year 1: (6,49)	0.13	−0.05	0.00	0.00	0.01	0.02	0.07	−0.01	−0.01	0.04	0.02	0.12	−0.06
43	No previous founding experience	−0.09	0.09	0.00	0.01	0.01	0.02	−0.23	0.12	−0.02	−0.03	−0.02	0.00	−0.01
44	Previously founded 1–3 firms	−0.10	0.03	0.00	0.01	0.03	0.00	−0.23	0.05	0.02	−0.01	0.01	−0.02	0.02
45	Previously founded more than 3 firms	0.10	−0.07	0.02	0.00	−0.02	0.00	0.19	−0.04	0.11	0.03	−0.02	0.01	0.03
46	Earlier invol. exit by one founder	0.00	0.00	−0.02	0.00	0.00	0.00	0.01	0.00	0.03	0.00	0.01	0.01	0.00
47	New patent	0.03	−0.02	0.00	0.00	−0.01	−0.01	0.05	−0.01	0.00	0.08	−0.01	−0.01	−0.02
48	Total assets year 1	0.13	−0.17	0.02	0.02	0.00	0.03	−0.01	−0.06	−0.07	0.03	−0.03	−0.02	0.02
49	Total assets year 1 is missing	−0.03	0.09	−0.01	−0.01	−0.01	−0.02	0.04	0.03	0.04	−0.02	0.02	0.02	−0.02
50	2nd quintile profits year 1	−0.05	−0.04	0.00	−0.01	0.00	0.00	0.00	0.00	0.00	0.00	−0.01	−0.01	0.01
51	3rd quintile profits year 1	−0.04	0.09	−0.01	−0.01	0.00	−0.02	−0.04	0.00	0.00	−0.01	−0.03	−0.02	0.02
52	4th quintile profits year 1	−0.02	−0.02	0.02	0.02	0.01	0.01	−0.06	0.00	−0.01	0.00	0.02	−0.02	0.00
53	5th quintile profits year 1	0.08	−0.09	0.02	0.01	0.00	0.04	−0.06	−0.04	−0.04	0.00	−0.05	−0.04	0.01
54	Previous exit at same address	0.05	−0.03	−0.12	−0.06	−0.06	−0.08	0.15	−0.05	−0.02	−0.02	−0.01	0.04	−0.02
55	>9 previous exits at same address	0.02	−0.02	−0.11	−0.04	−0.07	−0.08	0.13	−0.06	−0.03	−0.02	−0.03	0.00	−0.02
56	Address unshared	−0.06	0.04	0.04	0.04	0.03	0.05	−0.24	0.07	0.02	−0.01	−0.01	−0.01	−0.02
57	Address shared with 2–5 other firms	−0.06	0.01	0.05	0.05	0.08	0.06	−0.07	0.03	0.03	0.03	0.06	−0.01	0.00
58	Address shared with 6–10 other firms	0.04	−0.01	0.00	−0.02	−0.02	−0.02	0.08	−0.01	0.00	0.00	0.00	0.02	0.00
59	Address prev. used by 1–5 others	−0.07	0.02	0.07	0.05	0.07	0.07	−0.12	0.04	0.02	0.02	0.04	−0.03	0.00
60	Address prev. used by 6–10 others	0.01	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.02	0.01	0.02	−0.01
61	Address prev. used by 11–100 others	0.08	−0.02	−0.06	−0.05	−0.06	−0.06	0.14	−0.02	0.00	−0.01	−0.02	0.05	0.00
62	Address prev. used by > 100 others	0.06	−0.04	−0.10	−0.06	−0.09	−0.09	0.16	−0.07	−0.05	−0.02	−0.05	−0.01	0.00
63	Founded in 2013	0.02	−0.21	0.00	0.01	0.00	0.00	0.03	−0.02	−0.02	0.01	0.01	0.01	0.02
64	Founded in 2014	−0.05	0.38	−0.01	−0.01	0.00	−0.01	−0.05	0.02	0.03	−0.02	−0.02	−0.01	−0.01

(*continued on next page*)

(*continued*)

		1	2	3	4	5	6	7	8	9	10	11	12	13
65	BPS wordscore invol. exit	−0.01	0.09	−0.07	−0.02	0.05	−0.02	0.07	0.04	0.04	0.04	0.32	0.23	−0.23
66	Firm name index invol. exit	−0.06	0.11	−0.07	−0.03	0.02	−0.05	−0.22	0.27	0.13	0.00	0.06	0.06	0.01
67	Mean ind. perf. invol. exit	−0.02	0.00	−0.02	0.00	0.03	0.00	0.01	0.03	0.02	−0.06	0.42	0.29	−0.41
		1	2	3	4	5	6	7	8	9	10	11	12	13
1	Sector 5	1.00												
2	Sector is missing	−0.27	1.00											
3	In firm name: a proper danish word	−0.02	−0.04	1.00										
4	In firm name: Holding, capital, shares	−0.13	0.25	−0.15	1.00									
5	In firm name: female name	−0.02	0.01	0.11	0.03	1.00								
6	In firm name: male name	−0.02	0.03	0.13	0.10	0.16	1.00							
7	Eponymous firm name	−0.03	0.06	−0.01	0.29	0.06	0.29	1.00						
8	Short firm name	0.02	−0.04	−0.14	−0.16	−0.04	−0.09	−0.13	1.00					
9	Long firm name	−0.02	0.03	0.14	0.11	0.04	0.15	0.12	−0.21	1.00				
10	Firm name: w/geogr. location	−0.03	0.00	0.26	−0.02	−0.02	−0.02	−0.06	−0.09	0.10	1.00			
11	Firm name w/ Danmark, Danish, Dan	0.01	−0.02	0.08	−0.05	0.02	0.03	−0.04	−0.04	0.03	0.19	1.00		
12	LIX	0.08	0.01	−0.04	0.03	−0.01	−0.01	−0.01	0.00	0.00	−0.01	0.02	1.00	
13	Mean word length	0.04	0.03	−0.03	0.17	0.01	0.02	0.06	−0.02	0.03	0.00	−0.01	0.27	1.00
14	BPS lengths	0.05	0.02	−0.03	0.01	0.00	−0.02	−0.05	0.00	0.01	0.00	0.02	0.51	0.12
15	BPS 2nd quintile	−0.02	−0.02	0.02	0.00	0.00	0.01	0.02	0.00	0.00	0.01	0.00	−0.18	−0.02
16	BPS 3rd quintile	−0.01	−0.01	0.00	0.00	−0.01	0.00	0.00	−0.01	0.01	−0.01	−0.01	−0.04	0.05
17	BPS 4th quintile	−0.01	0.01	−0.02	0.05	−0.01	−0.01	0.01	0.00	0.00	−0.01	0.00	0.11	0.11
18	BPS 5th quintile	0.05	0.02	−0.03	0.01	0.00	−0.02	−0.04	0.00	0.00	0.00	0.02	0.45	0.08
19	Frequency of least common word	−0.09	0.14	−0.08	0.33	0.01	0.04	0.14	−0.04	0.02	−0.02	−0.04	−0.22	0.12
20	Frequency of most common word	−0.07	0.13	−0.10	0.28	0.00	0.00	0.06	−0.02	0.00	−0.02	−0.01	0.07	−0.10
21	Freq. least/freq. most common word	−0.02	0.03	0.00	0.08	0.01	0.03	0.07	−0.02	0.01	0.00	−0.02	−0.26	0.24
22	A geogr. term name is in BPS	−0.01	0.00	0.03	−0.05	0.01	0.00	−0.04	−0.02	0.03	0.08	0.02	0.04	−0.04
23	A male name is in BPS	0.01	−0.02	0.02	−0.05	0.00	0.01	−0.02	0.00	0.00	0.01	0.01	−0.04	−0.09
24	A female name is in BPS	0.01	−0.01	0.01	−0.03	0.05	0.02	−0.02	0.00	0.02	0.00	0.00	−0.01	−0.04
25	# employees year 1: 0	0.05	−0.09	0.02	−0.12	0.00	0.00	−0.03	0.02	−0.01	−0.01	0.00	0.00	−0.01
26	# employees year 1: 1	0.10	−0.12	0.03	−0.15	−0.01	0.00	−0.02	0.03	−0.03	−0.01	0.02	0.01	−0.01
27	# employees year 1: (2,4)	0.05	−0.11	0.06	−0.16	−0.01	0.00	−0.04	0.01	0.00	0.03	0.00	0.00	−0.02
28	# employees year 1: (5,9)	−0.01	−0.06	0.06	−0.11	0.00	0.01	−0.03	−0.01	0.01	0.04	0.00	−0.02	−0.02
29	# employees year 1: (6,49)	−0.01	−0.05	0.05	−0.10	0.00	0.01	−0.04	0.00	0.01	0.04	0.01	−0.01	0.01
30	No previous founding experience	−0.01	0.04	−0.04	0.11	0.03	0.05	0.15	0.00	0.00	−0.02	−0.03	−0.01	0.02
31	Previously founded 1–3 firms	0.01	−0.02	−0.02	0.08	−0.01	0.04	0.10	−0.01	−0.01	−0.03	−0.02	−0.01	0.03
32	Previously founded more than 3 firms	−0.01	−0.01	0.02	−0.08	−0.01	−0.05	−0.13	0.00	0.01	0.02	0.01	0.01	−0.02
33	Earlier invol. exit by one founder	0.00	−0.01	0.00	−0.04	−0.01	−0.02	−0.05	0.01	−0.01	0.01	0.01	−0.01	−0.02
34	New patent	0.05	−0.02	−0.02	−0.02	−0.01	−0.01	−0.02	0.01	−0.01	−0.01	0.00	0.03	0.01
35	Total assets year 1	−0.01	0.02	0.02	0.06	0.00	0.03	−0.02	−0.03	0.04	0.03	−0.01	0.04	0.04
36	Total assets year 1 is missing	−0.01	0.00	0.00	−0.06	0.00	−0.03	−0.02	0.01	−0.02	−0.01	0.01	−0.03	−0.04
37	2nd quintile profits year 1	0.00	0.00	−0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.01	−0.01
38	3rd quintile profits year 1	0.00	0.02	−0.02	0.02	−0.01	0.00	0.03	0.01	−0.01	−0.02	0.00	−0.01	0.00
39	4th quintile profits year 1	0.03	−0.03	0.02	−0.01	−0.01	0.02	0.03	0.00	0.00	0.00	−0.01	0.01	0.02

(*continued on next page*)

(*continued*)

		1	2	3	4	5	6	7	8	9	10	11	12	13
40	5th quintile profits year 1	−0.01	0.06	−0.01	0.13	0.01	0.03	0.01	−0.03	0.04	0.01	−0.01	0.03	0.06
41	Previous exit at same address	0.01	0.00	0.02	−0.11	−0.02	−0.05	−0.11	0.02	−0.02	−0.01	0.02	0.01	−0.02
42	>9 previous exits at same address	−0.01	0.05	0.05	−0.08	−0.01	−0.04	−0.06	0.01	−0.03	−0.02	0.01	0.01	0.00
43	Address unshared	−0.02	0.04	−0.02	0.10	0.01	0.05	0.13	−0.01	0.01	0.01	−0.02	0.00	0.02
44	Address shared with 2–5 other firms	0.00	−0.05	−0.01	0.04	−0.01	0.02	0.05	−0.01	0.00	−0.01	−0.01	−0.03	0.01
45	Address shared with 6–10 other firms	0.00	−0.01	−0.01	−0.02	0.01	−0.02	−0.04	0.01	0.00	0.00	0.01	0.01	−0.01
46	Address prev. used by 1–5 others	−0.01	−0.02	−0.01	0.06	−0.01	0.03	0.07	−0.01	0.01	0.00	−0.01	−0.01	0.01
47	Address prev. used by 6–10 others	0.00	−0.02	0.00	−0.03	0.00	−0.01	−0.04	0.01	−0.01	0.00	0.00	−0.01	−0.01
48	Address prev. used by 11–100 others	0.01	−0.02	0.00	−0.08	0.02	−0.03	−0.09	0.01	0.00	0.00	0.02	0.01	−0.02
49	Address prev. used by > 100 others	0.00	0.05	0.05	−0.09	−0.01	−0.06	−0.08	0.01	−0.02	−0.01	0.01	0.03	0.00
50	Founded in 2013	0.00	−0.03	0.01	0.01	0.01	0.01	−0.01	−0.01	0.01	0.01	0.00	0.00	0.01
51	Founded in 2014	0.00	0.04	−0.03	−0.01	−0.01	−0.01	0.02	0.02	−0.01	−0.02	0.00	0.01	−0.02
52	BPS wordscore invol. exit	0.02	−0.20	0.11	−0.43	0.00	−0.02	−0.14	0.08	−0.08	0.01	0.04	−0.13	−0.15
53	Firm name index invol. exit	0.00	−0.09	0.00	−0.08	0.02	−0.02	0.10	0.03	−0.06	−0.02	0.02	−0.04	−0.03
54	Mean ind. perf. invol. exit	−0.11	−0.09	0.07	−0.19	0.00	0.01	−0.07	0.04	−0.03	0.02	0.01	−0.09	−0.07
		1	2	3	4	5	6	7	8	9	10	11	12	13
1	BPS lengths	1.00												
2	BPS 2nd quintile	−0.27	1.00											
3	BPS 3rd quintile	−0.13	−0.25	1.00										
4	BPS 4th quintile	0.08	−0.25	−0.25	1.00									
5	BPS 5th quintile	0.72	−0.25	−0.26	−0.26	1.00								
6	Frequency of least common word	−0.44	0.16	0.05	−0.06	−0.39	1.00							
7	Frequency of most common word	0.17	−0.04	0.04	0.08	0.12	0.35	1.00						
8	Freq. least/freq. most common word	−0.43	0.15	−0.06	−0.20	−0.34	0.53	−0.36	1.00					
9	A geogr. term name is in BPS	0.19	−0.06	−0.03	0.00	0.16	−0.18	−0.02	−0.10	1.00				
10	A male name is in BPS	0.13	−0.03	−0.01	0.01	0.08	−0.15	−0.01	−0.10	0.05	1.00			
11	A female name is in BPS	0.13	−0.03	−0.02	0.00	0.10	−0.12	−0.03	−0.06	0.18	0.28	1.00		
12	# employees year 1: 0	−0.01	−0.01	−0.01	−0.01	0.00	−0.07	−0.08	0.00	−0.01	0.01	0.01	1.00	
13	# employees year 1: 1	−0.01	0.00	0.00	−0.01	0.00	−0.08	−0.08	−0.01	−0.02	0.01	0.00	−0.06	1.00
14	# employees year 1: (2,4)	−0.03	0.01	0.00	−0.02	−0.03	−0.08	−0.12	0.01	−0.02	0.01	0.00	−0.06	−0.07
15	# employees year 1: (5,9)	−0.03	0.02	−0.02	−0.02	−0.02	−0.06	−0.09	0.02	−0.01	0.00	0.00	−0.04	−0.05
16	# employees year 1: (6,49)	−0.02	0.00	−0.01	−0.02	0.00	−0.05	−0.09	0.02	0.03	0.00	0.04	−0.03	−0.04
17	No previous founding experience	−0.03	0.01	−0.01	−0.01	−0.02	0.06	0.02	0.03	−0.01	0.00	0.01	0.01	0.00
18	Previously founded 1–3 firms	−0.05	0.02	0.00	0.00	−0.04	0.05	0.00	0.04	−0.06	−0.01	−0.02	0.01	0.03
19	Previously founded more than 3 firms	0.04	−0.02	0.01	0.01	0.03	−0.03	0.01	−0.04	0.06	0.00	0.01	−0.02	−0.04
20	Earlier invol. exit by one founder	−0.01	0.00	0.00	0.00	0.00	−0.01	0.00	−0.01	0.01	0.00	0.00	−0.01	−0.02
21	New patent	0.04	−0.02	−0.01	0.01	0.04	−0.05	0.00	−0.03	−0.01	0.00	0.00	0.01	0.02
22	Total assets year 1	0.04	−0.01	0.00	0.03	0.02	0.03	0.04	−0.01	0.01	−0.01	−0.01	−0.01	0.07
23	Total assets year 1 is missing	−0.02	0.01	−0.01	−0.02	−0.01	−0.02	−0.03	0.01	0.00	0.01	0.01	0.00	−0.09
24	2nd quintile profits year 1	0.00	−0.01	0.01	0.01	0.00	0.00	0.02	−0.01	0.00	−0.01	−0.01	−0.01	0.00
25	3rd quintile profits year 1	0.00	0.01	0.00	−0.01	0.00	0.02	0.02	0.01	0.01	0.00	0.00	−0.01	0.00
26	4th quintile profits year 1	−0.02	0.01	0.01	0.00	−0.01	0.00	−0.03	0.01	−0.01	−0.01	−0.01	0.02	0.08

(*continued on next page*)

(*continued*)

		1	2	3	4	5	6	7	8	9	10	11	12	13
27	5th quintile profits year 1	0.03	−0.01	−0.01	0.02	0.02	0.06	0.06	0.01	0.00	−0.01	−0.01	−0.02	0.00
28	Previous exit at same address	0.03	−0.01	−0.01	0.00	0.02	−0.06	−0.03	−0.02	0.00	0.01	−0.01	0.01	−0.02
29	>9 previous exits at same address	0.04	−0.01	0.00	−0.01	0.03	−0.04	−0.01	−0.03	0.00	0.00	0.00	0.00	−0.01
30	Address unshared	−0.03	0.01	0.00	0.01	−0.02	0.05	0.03	0.02	−0.02	0.00	0.00	−0.01	0.01
31	Address shared with 2–5 other firms	−0.07	0.02	0.02	0.00	−0.05	0.03	−0.01	0.03	−0.04	0.00	−0.01	0.02	0.02
32	Address shared with 6–10 other firms	0.00	0.01	0.00	0.00	0.00	−0.01	0.00	−0.01	−0.01	0.00	0.00	−0.01	−0.02
33	Address prev. used by 1–5 others	−0.05	0.02	0.00	0.01	−0.04	0.04	0.01	0.03	−0.03	−0.02	−0.01	0.00	0.03
34	Address prev. used by 6–10 others	−0.01	0.00	0.00	0.00	−0.01	0.00	0.00	0.01	−0.01	0.01	0.00	0.00	−0.01
35	Address prev. used by 11–100 others	0.04	−0.01	−0.01	0.00	0.03	−0.05	−0.02	−0.02	0.02	0.01	0.00	0.00	−0.02
36	Address prev. used by > 100 others	0.08	−0.02	−0.01	−0.01	0.07	−0.06	−0.01	−0.05	0.06	0.00	0.01	0.00	−0.01
37	Founded in 2013	0.01	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.01	0.00	0.00	−0.01	0.00
38	Founded in 2014	−0.02	0.01	0.01	−0.01	−0.02	0.00	−0.01	0.00	−0.01	0.00	0.00	−0.01	−0.01
39	BPS wordscore invol. exit	−0.14	0.02	−0.05	−0.09	−0.10	−0.24	−0.33	0.04	−0.05	0.05	0.00	0.12	0.10
40	Firm name index invol. exit	−0.05	0.03	−0.01	−0.02	−0.04	−0.05	−0.09	0.03	−0.02	0.03	0.00	0.03	0.02
41	Mean ind. perf. invol. exit	−0.09	0.02	0.00	−0.04	−0.07	−0.09	−0.14	0.03	−0.02	0.03	0.00	0.07	0.05
		1	2	3	4	5	6	7	8	9	10	11	12	13
1	# employees year 1: (2,4)	1.00												
2	# employees year 1: (5,9)	−0.05	1.00											
3	# employees year 1: (6,49)	−0.04	−0.03	1.00										
4	No previous founding experience	−0.03	−0.02	−0.02	1.00									
5	Previously founded 1–3 firms	0.03	0.01	−0.03	−0.46	1.00								
6	Previously founded more than 3 firms	−0.01	0.00	0.01	−0.26	−0.48	1.00							
7	Earlier invol. exit by one founder	0.00	−0.01	0.00	−0.15	−0.04	0.30	1.00						
8	New patent	0.00	0.00	0.00	−0.02	−0.02	0.03	0.00	1.00					
9	Total assets year 1	0.10	0.09	0.09	−0.09	0.03	0.07	−0.02	0.01	1.00				
10	Total assets year 1 is missing	−0.09	−0.06	−0.05	0.06	−0.06	−0.03	0.02	−0.01	−0.90	1.00			
11	2nd quintile profits year 1	−0.03	−0.03	−0.04	−0.01	0.01	0.01	0.00	0.00	0.21	−0.34	1.00		
12	3rd quintile profits year 1	−0.02	−0.03	−0.04	0.01	0.02	−0.01	0.00	−0.01	0.16	−0.32	−0.11	1.00	
13	4th quintile profits year 1	0.09	0.03	0.01	−0.01	0.07	−0.02	−0.01	−0.01	0.31	−0.34	−0.11	−0.11	1.00
14	5th quintile profits year 1	0.01	0.04	0.06	−0.03	0.03	0.03	−0.02	−0.01	0.49	−0.34	−0.11	−0.11	−0.11
15	Previous exit at same address	0.01	0.02	0.03	−0.10	−0.08	0.08	0.11	0.01	−0.01	0.02	−0.02	−0.01	−0.03
16	>9 previous exits at same address	0.00	−0.01	0.02	−0.05	−0.07	0.02	0.02	0.01	−0.01	0.03	−0.01	0.00	−0.03
17	Address unshared	−0.02	−0.02	−0.02	0.26	−0.03	−0.12	−0.03	−0.02	−0.02	0.00	0.01	0.01	0.02
18	Address shared with 2–5 other firms	0.03	0.01	0.00	−0.05	0.19	−0.10	−0.02	0.00	−0.01	−0.01	0.01	0.01	0.04
19	Address shared with 6–10 other firms	0.00	0.01	0.00	−0.06	−0.04	0.08	0.01	0.00	0.02	0.00	−0.01	0.00	−0.01
20	Address prev. used by 1–5 others	0.01	0.00	−0.01	−0.04	0.19	−0.11	−0.03	−0.01	0.01	−0.03	0.02	0.01	0.04
21	Address prev. used by 6–10 others	0.00	0.01	0.01	−0.05	−0.01	0.06	0.02	0.01	0.02	0.00	−0.01	−0.01	0.01
22	Address prev. used by 11–100 others	0.01	0.02	0.01	−0.09	−0.10	0.14	0.04	0.01	0.02	0.01	−0.01	−0.01	−0.03

(*continued on next page*)

(*continued*)

		1	2	3	4	5	6	7	8	9	10	11	12	13
23	Address prev. used by >100 others	−0.01	−0.01	0.01	−0.07	−0.11	0.06	0.03	0.02	0.00	0.02	−0.01	0.00	−0.04
24	Founded in 2013	0.00	0.01	0.02	−0.03	0.00	0.02	0.00	0.01	0.06	−0.04	0.01	−0.01	0.00
25	Founded in 2014	−0.03	−0.02	−0.03	0.04	0.02	−0.01	0.02	−0.01	−0.09	0.04	−0.01	0.05	−0.01
26	BPS wordscore invol. exit	0.17	0.14	0.15	0.00	0.00	−0.06	0.02	−0.01	−0.14	0.10	−0.04	−0.03	−0.01
27	Firm name index invol. exit	0.04	0.01	0.00	0.12	0.04	−0.09	0.01	−0.01	−0.21	0.13	0.00	0.02	−0.03
28	Mean ind. perf. invol. exit	0.10	0.10	0.12	0.01	0.00	−0.05	0.01	−0.03	−0.06	0.05	−0.01	−0.03	−0.01

		1	2	3	4	5	6	7	8	9	10	11	12	13
1	5th quintile profits year 1	1.00												
2	Previous exit at same address	−0.01	1.00											
3	>9 previous exits at same address	−0.01	0.37	1.00										
4	Address unshared	−0.01	−0.20	−0.09	1.00									
5	Address shared with 2–5 other firms	−0.01	−0.29	−0.22	−0.42	1.00								
6	Address shared with 6–10 other firms	0.01	0.06	−0.08	−0.17	−0.39	1.00							
7	Address prev. used by 1–5 others	0.00	−0.36	−0.19	0.00	0.49	−0.20	1.00						
8	Address prev. used by 6–10 others	0.01	−0.01	−0.09	−0.13	0.09	0.20	−0.34	1.00					
9	Address prev. used by 11–100 others	0.00	0.36	−0.07	−0.22	−0.38	0.25	−0.47	−0.22	1.00				
10	Address prev. used by > 100 others	−0.01	0.42	0.69	−0.12	−0.29	−0.11	−0.25	−0.12	−0.16	1.00			
11	Founded in 2013	0.03	0.01	0.00	−0.01	−0.01	0.01	−0.01	0.00	0.00	0.02	1.00		
12	Founded in 2014	−0.04	−0.01	0.00	0.01	0.01	0.00	0.01	0.01	−0.01	−0.02	−0.55	1.00	
13	BPS wordscore invol. exit	−0.12	0.06	0.02	−0.01	0.04	0.00	0.01	0.02	0.01	−0.02	−0.03	0.03	1.00
14	Firm name index invol. exit	−0.18	0.01	−0.01	0.04	0.02	−0.02	0.01	0.00	0.00	−0.02	−0.03	0.05	0.19
15	Mean ind. perf. invol. exit	−0.06	0.03	−0.01	0.01	0.03	0.00	0.01	0.01	0.01	−0.04	−0.01	0.00	0.44

		1.00	2.00
1	Firm name index invol. exit	1.00	
2	Mean ind. perf. invol. exit	0.09	1.00

Appendix B. coefficient estimates

	Survival		High empl. growth		High RoA		Inno. subs. program		New patent	
	Coeff.	**m.e.**	**Coeff.**	**m.e.**	**Coeff.**	**m.e.**	**Coeff.**	**m.e.**	**Coeff.**	**m.e.**
Legal form: corporation	0.881***	0.718***	−0.258***	0.470***	0.589*	0.057	0.039	−0.028	0.0023	0.00032
Legal form: IVS	−0.989***	−0.435***	−0.382***	0.282	−0.281	−0.120	−0.015	−0.041	0.0013	−0.00010
At least one founders is firm	0.803***	−0.355***	0.408***	0.231*	1.129***	0.066	−0.014	0.051	0.0010	0.00057
At least one of founder has female first name	−0.072*	−0.123*	−0.01	0.197	0.772**	−0.007	−0.005	−0.001	0.0008	0.00042
Team	0.05	0.242***	−0.165***	0.555***	−0.698	0.004	0.011	−0.019	0.0028	−0.00022
In firm name: a proper danish word	−0.088**	0.134***	−0.037	−0.167	−0.484*	−0.008	0.005	−0.004	−0.0007	−0.00020
In firm name: Holding, capital, shares	−0.069	−1.272***	0.073	−1.173***	−0.639	−0.006	−0.042	0.009	−0.0038	−0.00023
In firm name: female name	0.003	−0.227***	−0.08	−0.775***	0.062	0.000	−0.008	−0.009	−0.0023	0.00003
In firm name: male name	0.079*	0.092	−0.014	−0.206	−0.03	0.007	0.004	−0.002	−0.0008	−0.00001
Eponymous firm name	0.222***	−0.098	0.077*	−0.221	0.735	0.018	−0.004	0.009	−0.0008	0.00040
Short firm name	−0.005	0.05	0.049	0.151	0.365	0.000	0.002	0.006	0.0006	0.00017
Long firm name	0.008	0.04	−0.028	−0.227	−0.763	0.001	0.002	−0.003	−0.0008	−0.00025
Firm name: w/geogr. location	0.222***	0.055	−0.091	−0.894***	−0.197	0.018	0.002	−0.011	−0.0025	−0.00007
Firm name w/Danmark, Danish, Dan	−0.281***	0.264**	0.148	0.174	−0.926	−0.028	0.012	0.019	0.0008	−0.00025
LIX	0.357***	0.227	0.084	0.311	0.499	0.031	0.009	0.010	0.0012	0.00021
Mean word length	−0.021**	−0.036***	−0.012	0.006	−0.093	−0.002	−0.001	−0.001	0.0000	−0.00004
BPS lengths	0.001	0.001	0.001	0.004***	0.004	0.000	0.000	0.000	0.0000	0.00000
BPS 2nd quintile	−0.035	0.018	0.024	−0.779***	0.331	−0.003	0.001	0.003	−0.0025	0.00015
BPS 3rd quintile	−0.018	−0.045	0.045	−0.21	0.279	−0.002	−0.002	0.005	−0.0008	0.00013
BPS 4th quintile	−0.036	0.011	0.077	−0.223	1.107*	−0.003	0.000	0.009	−0.0008	0.00066
BPS 5th quintile	−0.048	−0.043	0.175*	−0.119	1.228*	−0.004	−0.002	0.022	−0.0005	0.00078
Frequency of least common word	0.009	0.014	0.007	0.017	−0.181**	0.001	0.001	0.001	0.0001	−0.00007
Frequency of most common word	−0.02	0.028	−0.035*	0.113**	0.265**	−0.002	0.001	−0.004	0.0005	0.00011
Freq. least/freq. most common word	0.002	0	−0.145*	−0.022	2.323***	0.000	0.000	−0.017	−0.0001	0.00095
A geogr. term name is in BPS	0.007	−0.457***	0.055	−0.133	–	0.001	−0.015	0.007	−0.0005	–
A male name is in BPS	0.235**	0.099	0.226*	−0.218	−1.566	0.019	0.004	0.029	−0.0008	−0.00033
A female name is in BPS	−0.169	0.031	−0.189	−0.231	–	−0.016	0.001	−0.021	−0.0008	–
# employees year 1: 0	−0.142*	1.055***	0.196**	0.355*	0.49	−0.013	0.067	0.025	0.0017	0.00025
# employees year 1: 1	0.276***	−0.670***	0.396***	0.564***	0.391	0.022	−0.021	0.054	0.0029	0.00019
# employees year 1: (2,4)	−0.286***	−0.922***	0.430***	0.780***	0.581	−0.028	−0.026	0.059	0.0045	0.00031
# employees year 1: (5,9)	−0.261***	−1.276***	0.288***	1.247***	0.514	−0.025	−0.031	0.038	0.0095	0.00027
# employees year 1: (6,49)	−0.319***	−3.176***	0.535***	1.431***	−0.948	−0.032	−0.043	0.077	0.0121	−0.00026
No previous founding experience	1.902***	0.878***	1.180***	0.298	0.343	0.114	0.046	0.181	0.0013	0.00016
Previously founded 1–3 firms	1.985***	1.072***	1.026***	0.355**	0.564	0.178	0.045	0.127	0.0014	0.00024
Previously founded more than 3 firms	2.563***	1.242***	0.929***	0.583***	1.250***	0.143	0.069	0.133	0.0028	0.00079
Earlier invol. exit by one founder	−0.574***	0.081	0.048	0.103	−0.323	−0.061	0.003	0.006	0.0004	−0.00012
Patent at start	0.349	0.659***	−1.232***	0.986***	3.214***	0.027	0.036	−0.095	0.0067	0.00960
Total assets year 1	0.141***	0.079***	−0.192***	0.046	0.12	0.012	0.003	−0.023	0.0002	0.00005
Total assets year 1 is missing	0.362***	0.917***	−0.930***	−0.105	1.373	0.032	0.039	−0.110	−0.0004	0.00061
2nd quintile profits year 1	0.498***	−0.347***	−0.160**	−0.274	0.526	0.038	−0.012	−0.018	−0.0010	0.00027
3rd quintile profits year 1	0.726***	−0.022	−0.363***	−0.588**	0.499	0.051	−0.001	−0.039	−0.0019	0.00025
4th quintile profits year 1	0.812***	0.077	0.503***	−0.115	0.636	0.056	0.003	0.069	−0.0004	0.00034
5th quintile profits year 1	1.273***	0.236**	0.938***	−0.454**	−1.501*	0.077	0.010	0.144	−0.0015	−0.00037
Previous exit at same address	−0.240***	−0.004	0.072	−0.034	−0.159	−0.022	0.000	0.009	−0.0001	−0.00006
>9 previous exits at same address	−0.284***	−0.063	0.14	−0.262	−0.022	−0.028	−0.002	0.018	−0.0009	−0.00001
Address unshared	0.011	0.056	0.175**	−0.403	0.26	0.001	0.002	0.022	−0.0014	0.00012

(*continued on next page*)

(*continued*)

	Survival		High empl. growth		High RoA		Inno. subs. program		New patent	
	Coeff.	**m.e.**	**Coeff.**	**m.e.**	**Coeff.**	**m.e.**	**Coeff.**	**m.e.**	**Coeff.**	**m.e.**
Address shared with 2–5 other firms	−0.139*	−0.091	0.056	−0.142	0.006	−0.012	−0.004	0.007	−0.0006	0.00000
Address shared with 6–10 other firms	−0.275***	−0.063	0.02	0.194	0.167	−0.026	−0.002	0.002	0.0008	0.00007
Address previously used by 1–5 others	−0.067	−0.065	0.008	−0.307*	0.014	−0.006	−0.003	0.001	−0.0012	0.00001
Address previously used by 6–10 others	−0.08	0.056	−0.072	−0.614***	0.305	−0.007	0.002	−0.008	−0.0020	0.00014
Address previously used by 11–100 others	0.123	−0.046	0.053	−0.582**	0.409	0.011	−0.002	0.006	−0.0020	0.00019
Address previously used by > 100 others	0.427***	0.031	−0.119	−0.105	−0.052	0.033	0.001	−0.014	−0.0004	−0.00002
Mean ind. perf. index	7.421***	20.018***	9.653***	35.820***	28.396***	0.655	0.796	1.158	0.1434	0.01167
Firm name index	5.361***	13.830***	7.357***	35.044***	42.233***	0.473	0.550	0.883	0.1403	0.01735
BPS wordscore invol. exit	4.206***	7.409***	6.088***	6.748***	5.085***	0.371	0.295	0.730	0.0270	0.00209

The models additionally include sets of year, region and sector dummies as well as constant terms. Robust standard errors. "m.e." refers to the corresponding marginal effect. The asteriks' ***, ** and * denote marginal significance at the one, five and ten percent level.

Appendix C. robustness checks — reduced sets of variables

AUC full model	**AUC # empl.**		**AUC Pa-tents**		AUC **Finan-cial vars.**		AUC **BPS word-scores**		AUC **Mean ind. index**		AUC **Founder name index**		**AUC All indices**	
	excl.	**Δ**	**excl.**	**Δ**	**excl.**	**Δ**	**excl.**	**Δ**	**excl.**	**Δ**	**excl.**	**Δ**	**excl.**	**Δ**
Survival														
0.82	0.82	−0.03	0.82	0.01	0.82	−0.10	0.80	−2.79	0.81	−1.93	0.83	0.18	0.80	−2.38
High employment growth														
0.82	0.79	−3.48	0.82	−0.08	0.82	0.37	0.82	−0.22	0.78	−5.23	0.83	0.66	0.78	−4.54
High return on assets														
0.68	0.68	0.06	0.68	0.06	0.69	1.15	0.66	−2.24	0.65	−3.91	0.68	0.39	0.66	−3.36
Innovation subsidy program														
0.90	0.90	−0.36	0.90	−0.11	0.91	0.28	0.90	−0.07	0.84	−6.57	0.91	0.43	0.84	−6.59
New patent														
0.84	0.83	−0.83	0.84	0.16	0.84	−0.11	0.86	2.61	0.80	−5.08	0.84	−0.48	0.80	−4.88

The "base" set of variables is included in all specifications. All changes Δ refer to the percentage differences relative to the full model.

Author's statement

Ulrich Kaiser: initial idea, estimation, writing, Johan Kuhn: initial idea, data download, data preparation, first estimation

References

Åstebro, T., Winter, J.K., 2012. More than a dummy: the probability of failure, survival and acquisition of firms in financial distress. Eur. Manag. Rev. 9 (1), 1–17.

Agrawal, Vineet, Taffler, Richard, 2008. Comparing the performance of market-based and accounting-based bankruptcy prediction models. J. Bank Finan. 32, 1541–1551.

Ahuja, G., Coff, R.W., Lee, P.M., 2005. Managerial foresight and attempted rent appropriation: insider trading on knowledge of imminent breakthroughs. Strat. Manag. J. 26 (9), 791–808.

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Finance 23, 589609.

Altman, E.I., 1984. The success of business failure prediction models: an international survey. J. Bank. Finance 8, 171–184.

Amit, R., Glosten, L., Muller, E., 1990. Entrepreneurial ability, venture investments, and risk sharing. Manag. Sci. 36 (10), 1232–1245.

Antweiler, W., Frank, M.Z., 2004. Is all that talk just Noise? The information content of Internet stock Message boards. J. Finance 59 (3), 125994.

Arora, A., Nandkumar, A., 2011. Cash-Out or flameout! Opportunity cost and entrepreneurial strategy: theory, and evidence from the information security industry. Manag. Sci. 57 (10), 1844–1860.

Arundel, A., Kabla, I., 1998. What percentage of innovations are patented? Empirical estimates for European firms. Res. Pol. 27, 127–141.

Audretsch, D.B., Mahmood, T., 1995. New firm survival: new results using a hazard function. Rev. Econ. Stat. 77 (1), 97–103.

Audretsch, D., Santarelli, E., Vivarelli, M., 1999. Start-up size and industrial dynamics: some evidence from Italian manufacturing. Int. J. Ind. Organ. 17, 965–983.

Banbura, M., Giannone, D., Modugno, M., Reichlin, L., 2013. Now-Cast-ing and the real-time data Flow. In: Timmermann, A., Elliot, G. (Eds.), Handbook of Economic Forecasting, 2A. Sebastopol: OReilly Media, 195237.

Barney, J.B., 2001. Gaining and Sustaining Competitive Advantage, second ed. Prentice-Hall, Upper Saddle River (NJ).

Baron, R.A., Ensley, M.D., 2006. Opportunity recognition as the detection of meaningful patterns: evidence from comparisons of novice and experienced entrepreneurs. Manag. Sci. 52 (9), 1331–1344.
Bates, T., 2005. Analysis of young, small firms that have closed: delineating successful from unsuccessful closures. J. Bus. Ventur. 20 (3), 343–358.
Baum, J.R., Wally, S., 2003. Strategic decision speed and firm performance. Strat. Manag. J. 24 (11), 1107129.
Belsley, D., Kuh, E., Welsch, R., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley & Sons.
Blanchflower, D.G., Oswald, A.J., 1998. What makes an entrepreneur? J. Labor Econ. 16 (1), 2660.
Brush, C.G., Vanderwerf, P.A., 1992. A comparison of methods and sources for obtaining estimates of new venture performance. J. Bus. Ventur. 7, 157170.
Björnsson, C.H., 1968. Läsbarhet. Liber, Stockholm.
Blundell, R., Griffith, R., van Reenen, J., 1995. Dynamic count data models of technological innovation. Econ. J. 105, 333–344.
Bonardo, D., Paleari, S., Vismara, S., 2011. Valuing university-based firms: the effects of academic affiliation on IPO performance. Enterpren. Theor. Pract. 35 (4), 755–776.
Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30 (7), 1145–1159.
Brüderl, J., Preisendörfer, P., Ziegler, R., 1992. Survival chances of newly founded business organizations. Am. Socio. Rev. 4 (1), 227–242.
Card, David, Chetty, Raj, Feldstein, Martin, Saez, Emmanuel, et al., 2010. Expanding access to administrative data for research in the United States working paper. https://eml.berkeley.edu/~saez/card-chetty-feldstein-saezNSF10dataaccess.pdf.
Cassar, G., 2014. Industry and startup experience on entrepreneur forecast performance in new firms. J. Bus. Ventur. 29, 137–151.
Catalini, C., Stern, J. Guzman S., 2019. Passive versus Active Growth: Evidence from Founder Choices and Venture Capital Investment. NBER Working. Paper No. w26073.
Carroll, G.R., 1993. A sociological view on why firms differ. Strat. Manag. J. 14, 237–249.
Carrol, G.R., Hannan, M.T., 2000. The Demography of Corporation and Industries. Princeton University Press, Princeton (NJ).
Chandler, G.N., Hanks, S.H., 1993. Measuring the performance of emerging businesses: a validation study. J. Bus. Ventur. 8, 391408.
Chava, S., Jarrow, R.A., 2004. Bankruptcy prediction with industry effects. Rev. Finance 8, 537–569.
Clarysse, B., Tartari, V., Salter, A., 2011. The impact of entrepreneurial capacity, experience and organizational support on academic entrepreneurship. Res. Pol. 40 (8), 10841093.
Cooper, A.C., 1993. Challenges in predicting new firm performance. J. Bus. Ventur. 8, 241–253.
Cooper, A.C., Gimeno-Gascon, F.J., Woo, C.Y., 1994. Initial human and financial capital as predictors of new venture performance. J. Bus. Ventur. 9, 371395.
Cope, J., 2011. Entrepreneurial learning from failure: an interpretive phenomenological analysis. J. Bus. Ventur. 26 (6), 604–623.
Corea, F., 2018. An Introduction to Data. Springer.
Cornett, M.M., Tehranian, H., 1992. Changes in corporate performance associated with bank acquisitions. J. Financ. Econ. 31 (2), 211–234.
Dambolena, I.G., Khoury, S.J., 1980. Ratio stability and corporate failure. J. Finance 35 (4), 1017–1026.
Davidsson, P., Honig, B., 2003. The role of social and human capital among nascent entrepreneurs. J. Bus. Ventur. 18 (3), 301–331.
Davis, S.J., Haltiwanger, J., Schuh, S., 1996. Small business and job creation: dissecting the myth and reassessing the facts. Small Bus. Econ. 8, 297315.
Delmar, F., Davidsson, P., Gartner, W.B., 2003. Arriving at the high-growth firm. Jounral of Business Venturing 18, 189–216.
Delmar, F., Shane, S., 2004. Legitimating first: organizing activities and the survival of new ventures. J. Bus. Ventur. 19, 385–410.
Denrell, J., 2004. Random walks and sustained competitive advantage. Manag. Sci. 50 (7), 922–934.
Denrell, J., Fang, C., Liu, C., 2015. Perspective-chance Explanations in the management sciences. Organ. Sci. 26 (3), 923–940.
Detienne, D.R., Wennberg, K., 2014. What do we really mean when we talk about "exit"? — a critical review of research on entrepreneurial exit. Int. Small Bus. J. 32 (1), 4–16.
Dencker, J.C., Gruber, M., 2015. The effects of opportunities and founder experience on new firm performance. Strat. Manag. J. 36, 1035–1052.
Diffey, C., 2019. Motherbrain: How AI Is Helping This VC Firm to Pick the Next Big Startup. TechRound, May 6, 2019. https://techround.co.uk/news/motherbrain-using-ai-to-pick-start-up/.
Eesley, C.E., Hsu, D.H., Roberts, E.B., 2014. The contingent effects of top management teams on venture performance: aligning founding team composition with innovation strategy and commercialization environment. Strat. Manag. J. 35 (12), 1798–1817.
Einav, L., Levin, J., 2013. The Data Revolution and Economic Analysis. SIEPR Discussion Paper 12-017.
Eisenhardt, K.M., 1989. Making fast strategic decisions in high-velocity environments. Acad. Manag. J. 32 (3), 543576.
Eisenhardt, K.M., Schoonhoven, C.B., 1990. Organizational growth: Linking founding team, strategy, environment, and GrowthAmong U.S. Semiconductor ventures, 1978-1988. Administratice Science Quarterly 35 (3), 504–529.
Ensley, M.D., Hmieleski, K.M., 2005. A comparative study of new venture top management team composition, dynamics and performance between university-based and independent startups. Res. Pol. 34, 10911105.
Fryer, R.G., Levitt, S.D., 2004. The causes and consequences of distinctively black names. Q. J. Econ. 119 (3), 767805.
Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/hierarchical Models. Cambridge University Press.
Gentzkow, M., Kelly, B.T., Taddy, M., 2019. Text as data. J. Econ. Lit. 57 (3), 535–574.
Gepp, A., Kumar, K., Bhattacharya, S., 2010. Business failure prediction using decision trees. J. Forecast. 29, 536–555.
Gerhards, J., Hans, S., 2009. From hasan to Herbert: NameGiving patterns of Immigrant Parents between Acculturation and ethnic Maintenance. Am. J. Sociol. 114 (4), 1102–1128.
Geroski, P., 1999. The Growth of Firms in Theory and Practice. Centre for Economic Policy Research working paper 2092.
Geroski, P., 2005. Understanding the implication of empirical work on corporate growth rates. Manag. Decis. Econ. 26 (2), 129–138.
Ghassemi, M.M., Song, C., Alhanai, T., 2020. The Automated Venture Capitalist: Data and Methods to Predict the Fate of Startup Ventures. Association for the Advancement of ArtificialIntelligence.
Gibrat, R., 1931. Les inegalites economiques; applications: aux inegalite's des richesses, a la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel. Librairie du Recueil Sirey, Paris.
Gimmon, E., Levie, J., 2010. Founders human capital, external investment, and the survival of new high-technology ventures. Res. Pol. 39, 1214–1226.
Goldstein, J.R., Stecklov, G., 2016. From Patrick to John F.: ethnic names and occupational success in the last era of mass migration. Am. Socio. Rev. 81 (1), 85106.
Gompers, P., Kovner, A.R., Lerner, J., Scharfstein, D., 2006. Skill vs. luck in entrepreneurship and venture capital: evidence from serial entrepreneurs. J. Financ. Econ. 96, 1832.
Greenberg, J., Mollick, E.R., 2018. Sole Survivors: Solo Ventures versus Founding Teams. https://doi.org/10.2139/ssrn.3107898. NYU working paper.
Greene, W., 2017. Econometric Analysis. Pearson.
Griliches, Z., 1990. Patent statistics as economic indicators: a survey. J. Econ. Lit. 28 (4), 1661–1990.
Guzman, J., Stern, S., 2015. Where is Silicon Valley? Science 347 (6222), 606–609.
Guzman, J., Stern, S., 2016. The State of American Entrepreneurship: New Estimates of the Quantity and Quality of Entrepreneurship for 32 US States, 1988-2014. NBER Working Paper 22095.
Hand, D.J., 2001. Measuring diagnostic accuracy of statistical prediction model. Stat. Neerl. 55, 3–16.
Hannan, M.T., Carroll, R.G., 1992. The population ecology of organizations. Am. Socio. Rev. 82, 929–964.
Hannan, M.T., Freeman, J., 1977. Structural inertia and organizational change. Am. Socio. Rev. 49, 149–164.
Hannan, M.T., Freeman, J., 1989. Organizational Ecology. Harvard University Press, Cambridge.
Harrison, J.R., 2004. Models of growth in organizational ecology: a simulation assessment. Ind. Corp. Change 13 (1), 243–261.
Hayward, M., Forster, W., Fredrickson, B., 2010. Beyond hubris: how highly confident entrepreneurs rebound to venture again. J. Bus. Ventur. 25 (6), 569–578.
Headd, B., 2003. Redefining business success: Distinguishing between closure and failure. Small Bus. Econ. 21 (1), 5161.

Henderson, A.D., Raynor, M.E., Ahmed, M., 2012. How long must a firm be great to rule out chance? Benchmarking sustained superior performance without being fooled by randomness. Strat. Manag. J. 33, 387–406.
Hopenhayn, H.A., 1992. Entry, exit, and firm dynamics in long run equilibrium. Econometrica 60 (5), 1127–1150.
Huyghebaert, N., Gaeremynck, A., Roodhooft, F., van de Gucht, L.M., 2000. New firm survival: the effect of start-up characteristics. J. Bus. Finance Account. 27 (5), 627–651.
Jovanovic, B., 1982. Selection and the evolution of industry. Econometrica 50 (3), 649670.
Kaiser, U., Kongsted, H.C., Rønde, T., 2015. Does the mobility of R&D labor increase innovation? J. Econ. Behav. Organ. 110, 91–105.
Kaiser, U., Kongsted, H.C., Laursen, K., Ejsing, A.-K., 2018. Experience matters: the role of academic scientist mobility for industrial innovation. Strat. Manag. J. 39 (7), 1935–1958.
Keats, B.W., 1988. The vertical construct validity of business economic performance measures. J. Appl. Behav. Sci. 24, 151160.
Kirsch, D., Goldfarb, B., Gera, A., 2009. Form or substance: the role of business plans in venture capital decision making. Strat. Manag. J. 30, 487515.
Krishna, A., Agrawal, A., Choudhary, A., 2016. Predicting the outcome of startups: less failure, more success. IEEE 16th International Conference on Data Mining Workshops (ICDMW) 798–805.
Laitinen, E., 1992. Prediction of failure of a newly founded firm. J. Bus. Ventur. 7, 323–340.
Laver, M., Benoit, K., Garry, J., 2003. Extracting policy positions from political texts using words as data. Am. Polit. Sci. Rev. 97 (2), 311–331.
Lubatkin, M., Shrieves, R.E., 1986. Towards reconciliation of market performance measures to strategic management research. Acad. Manag. Rev. 11, 497512.
March, J.G., Sutton, R.I., 1997. Organizational performance as a dependent variable. Organ. Sci. 8, 698706.
Mata, J., Portugal, P., 1994. Life duration of new firms. J. Ind. Econ. 42 (3), 227245.
McFadden, D., 1973. Conditional logit analysis of Qualitative choice Behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York.
Mehrabian, A., 1997. Impressions created by given names. Names 45 (1), 19–33.
Miller, C.C., Washburn, N.T., Glick, W.H., 2013. Perspective — the myth of firm performance. Organ. Sci. 24 (3), 948–964.
Mintzberg, H., 1990. The design school: Reconsidering the basic premises of strategic management. Strat. Manag. J. 11 (3), 171195.
Morgan, N.A., Vorhies, D.W., Mason, C.H., 2009. Market orientation, marketing capabilities, and firm performance. Strat. Manag. J. 30 (8), 909–920.
Nelson, F., Winter, S., 1982. An Evolutionary Theory of Economic Change. Harvard University Press, Cambrigde MA.
Nielsen, K., Sarasvathy, S.D., 2016. A market for lemons in serial entrepreneurship? Exploring type I and type II errors in the restart decision. Acad. Manag. Discov. 2 (3), 247–271.
Nikiforou, A., Dencker, J.C., Gruber, M., 2019. Necessity entrepreneurship and industry choice in new firm creation. Strat. Manag. J. 40 (13), 2165–2190.
Oriani, R., Sobrero, M., 2008. Uncertainty and the market Valuation of R&D within a real Options logic. Strat. Manag. J. 29 (4), 343–361.
Palmer, M., 2017. Artificial intelligence is guiding venture capital to startups. Financ. Times. https://www.ft.com/content/dd7fa798-bfcd-11e7-823b-ed31693349d3. Dec. 11, 2017.
Parker, S.C., 2008. Ch. 2: statistical Issues in applied entrepreneurship research: data, methods and Challenges. In: Congregado, E. (Ed.), Measuring Entrepreneurship: Building a Statistical System. Springer, Boston (MA), pp. 9–20.
Pennings, L., Lee, K., van Witteloostuijn, A., 1998. HumanCapital, social capital, and firm dissolution. Acad. Manag. J. 425440.
Penrose, E., 1959. The Theory of the Growth of the Firm. Oxford Univ. Press, Oxford.
Plehn-Dujowich, J., 2010. A theory of serial entrepreneurship. Small Bus. Econ. 35 (4), 377–398.
Porter, M.E., 1980. Competitive Strategy: Techniques for Analyzing Industries and Competitors. Free Press, New York.
Schendel, D.E., Hofer, C.W., 1979. Strategic Management: A New View of Business Policy and Planning. Little-Brown, Boston).
Scott, W.R., 1987. Organizations: Rational, Natural and Open Systems. Prentice Hall, Englewood Cliffs, NJ.
Simons, D.J., Shoda, Y., Lindsay, D.S., 2017. Constraints on generality (COG): a proposed addition to all empirical papers. Perspective on Psychological Science 12 (6), 1123–1128.
Stinchcombe, A.L., 1965. Social structure and organization. In: March, J.G. (Ed.), Handbook of Organizations. Rand McNally, Chicago, p. 142193.
Sutton, J., 1997. Gibtrats legacy. J. Econ. Lit. 35, 40–59.
Taddy, M., 2013. Rejoinder: efficiency and structure in MNIR. J. Am. Stat. Assoc. 108 (503), 77274.
Tetlock, P.C., 2007. Giving content to investor Sentiment: the role of Media in the stock market. J. Finance 62 (3), 113968.
Uhlbach, W.H., Tartari, V., Kongsted, H.C., 2019. Beyond Scientific Excellence: Are Internationally Mobile Researchers More Likely to Become Academic Entrepreneurs. Copenhagen Business School mimeo.
van Praag, C.M., 2003. Business survival and success of young small business Owners: an empirical analysis. Small Bus. Econ. 21, 1–17.
Visintin, F., Pittino, D., 2014. Founding team composition and early performance of university-based spin-o companies. Technovation 34, 31–43.
Wagner, J., 1992. Firm size, firm growth and persistence of chance. Testing GIBRATs law with establishment data from Lower Saxony, 19781989. Small Bus. Econ. 4, 125131.
Wagner, J., 2002. Taking a Second Chance: Entrepreneurial Restarters in Germany. The Institute for the Study of Labor (IZA) Discussion Paper Series.
Wang, G., Ma, J., Yang, S., 2014. An improved boosting based on feature selection for corporate bankruptcy prediction. Expert Syst. Appl. 41, 2353–2361.
Wennberg, K., Wiklund, J., Wright, M., 2011. The eectiveness of university knowledge spillovers: performance dierences between university spinos and corporate spinos. Res. Pol. 40, 1128–1143.
Westhead, P., Ucbasaran, D., Wright, M., Binks, M., 2005. Novice, serial and portfolio entrepreneur behaviour and contributions. Small Bus. Econ. 25, 109–132.
Wooldridge, J.M., 2010. Econometric Analysis of Cross Section and Panel Data. MIT press.
Zahra, S.A., van de Velde, E., Larraneta, B., 2007. Knowledge conversion capability and the performance of corporate and university spin-os. Ind. Corp. Change 16 (4), 569–608.